

10/500112

PCT/JP03/00011

10 2003 PCT/JP 25 JUN 2003

日本国特許 10序
JAPAN PATENT OFFICE

06.01.03

別紙添付の書類に記載されている事項は下記の出願書類に記載されて
いる事項と同一であることを証明する。

This is to certify that the annexed is a true copy of the following application as filed
with this Office

出願年月日

Date of Application:

2001年12月28日

REC'D 03 MAR 2003

WIPO PCT

出願番号

Application Number:

特願2001-402081

[ST.10/C]:

[JP2001-402081]

出願人

Applicant(s):

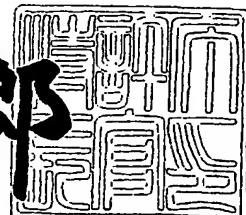
セレスター・レキシコ・サイエンシズ株式会社

PRIORITY DOCUMENT
SUBMITTED OR TRANSMITTED IN
COMPLIANCE WITH
RULE 17.1(a) OR (b)

2003年 2月12日

特許庁長官
Commissioner,
Japan Patent Office

太田信一郎



出証番号 出証特2003-3006524

【書類名】 特許願
 【整理番号】 PCLA-13299
 【提出日】 平成13年12月28日
 【あて先】 特許庁長官殿
 【国際特許分類】 G06F 17/30
 C07K 1/00

【発明者】

【住所又は居所】 千葉県千葉市美浜区中瀬1丁目3番地 幕張テクノガーデンD17 セレスター・レキシコ・サイエンシズ株式会社内

【氏名】 上村 泰央

【発明者】

【住所又は居所】 千葉県千葉市美浜区中瀬1丁目3番地 幕張テクノガーデンD17 セレスター・レキシコ・サイエンシズ株式会社内

【氏名】 蓬萊 尚幸

【特許出願人】

【識別番号】 500520628

【氏名又は名称】 セレスター・レキシコ・サイエンシズ株式会社

【代理人】

【識別番号】 100089118

【弁理士】

【氏名又は名称】 酒井 宏明

【選任した代理人】

【識別番号】 100113103

【弁理士】

【氏名又は名称】 香島 拓也

【手数料の表示】

【予納台帳番号】 036711

【納付金額】 21,000円

【提出物件の目録】

【物件名】 明細書 1

【物件名】 図面 1

【物件名】 要約書 1

【包括委任状番号】 0117745

【ブルーフの要否】 要

【書類名】 明細書

【発明の名称】 RNA配列解析装置、RNA配列解析方法、プログラム、および、記録媒体

【特許請求の範囲】

【請求項1】 RNA二次構造の構造トポロジーと、当該構造トポロジーに適合する生成文法とを対応付けて格納する文法格納手段と、

RNA配列を上記生成文法に適用して構文解析木を導出する構文解析手段と、上記構文解析手段にて導出された上記構文解析木に対して適合度の計算を行う適合度計算手段と、

上記適合度計算手段により計算された上記適合度のうち予め定めた条件を満たす適合度である上記構文解析木を上記適合度が高い順にソートするソート手段と

上記ソート手段によりソートされた上記構文解析木を上記RNA配列の二次構造の候補として出力する出力手段と、

を備えたことを特徴とするRNA配列解析装置。

【請求項2】 RNA二次構造の構造トポロジーと、当該構造トポロジーに適合する生成文法とを対応付けて格納する文法格納手段と、

RNA配列を上記生成文法に適用して構文解析木を導出する構文解析手段と、上記構文解析手段にて導出された上記構文解析木に対して適合度の計算を行う適合度計算手段と、

上記適合度計算手段により計算された上記適合度のうち予め定めた条件を満たす適合度である上記構文解析木を導出した上記RNA配列を上記構造トポロジーの二次構造を持つRNA配列の候補として出力する出力手段と、

を備えたことを特徴とするRNA配列解析装置。

【請求項3】 RNA二次構造の構造トポロジーと、当該構造トポロジーに適合する生成文法とを対応付けて格納する文法格納手段と、

RNA配列を上記生成文法に適用して構文解析木を導出する構文解析手段と、上記構文解析手段にて導出された上記構文解析木に対して適合度の計算を行う適合度計算手段と、

上記適合度計算手段により計算された上記適合度のうち予め定めた条件を満たす適合度である上記構文解析木を導出した上記RNA配列を抽出する抽出手段と

上記構造トポロジーと上記RNA配列とを二次元マトリックスで表示し、上記二次元マトリックスにおいて上記抽出手段にて抽出された上記RNA配列と上記構造トポロジーに対応する格子部分をマークすることにより、上記RNA配列間で共通に有する構造トポロジーを可視化する共通構造マトリックス作成手段と、を備えたことを特徴とするRNA配列解析装置。

【請求項4】 RNA二次構造の構造トポロジーと、当該構造トポロジーに適合する生成文法とを対応付けて格納する文法格納手段と、

利用者が入力したDNA配列から転写されるRNA配列を作成するRNA配列作成手段と、

上記RNA配列作成手段により作成された上記RNA配列に対して上記生成文法を適用して構文解析木を導出する構文解析手段と、

上記構文解析手段にて導出された上記構文解析木に対して適合度の計算を行う適合度計算手段と、

上記適合度計算手段により計算された上記適合度のうち予め定めた条件を満たす適合度である上記構文解析木を導出した上記RNA配列に対応する上記DNA配列部分を遺伝子の候補として予測する遺伝子予測手段と、

を備えたことを特徴とするRNA配列解析装置。

【請求項5】 RNA二次構造の構造トポロジーと、当該構造トポロジーに適合する生成文法とを対応付けて格納する文法格納手段と、

RNA配列を上記生成文法に適用して構文解析木を導出する構文解析手段と、

上記構文解析手段にて導出された上記構文解析木に対して適合度の計算を行う適合度計算手段と、

上記適合度計算手段により計算された上記適合度に基づいて上記RNA配列間の類似度を計算する類似度計算手段と、

を備えたことを特徴とするRNA配列解析装置。

【請求項6】 RNA二次構造の構造トポロジーと、当該構造トポロジーに

適合する生成文法とを対応付けて格納する文法格納手段と、

RNA配列を上記生成文法に適用して構文解析木を導出する構文解析手段と、

上記構文解析手段にて導出された上記構文解析木に対して適合度の計算を行う

適合度計算手段と、

上記適合度計算手段により計算された上記適合度のうち予め定めた条件を満たす適合度である上記構文解析木を導出した上記RNA配列を抽出する抽出手段と

上記構造トポロジーと上記RNA配列とを二次元マトリックスで表示し、上記二次元マトリックスにおいて上記抽出手段にて抽出された上記RNA配列と上記構造トポロジーに対応する格子部分に上記適合度を表示する適合度マトリックスを作成する適合度マトリックス作成手段と、

上記適合度マトリックス作成手段にて作成された上記適合度マトリックスについて、上記適合度により上記構造トポロジーをソートし、他のRNA配列について当該ソートされた上記構造トポロジーの順番に対応する上記生成文法により構文解析を行い上記適合度が最大となる上記構文解析木を求め、予め定めた条件を満たす上記適合度を持つ上記構文解析木に対応する上記他のRNA配列を抽出する共通構造抽出手段と、

を備えたことを特徴とするRNA配列解析装置。

【請求項7】 RNA二次構造の構造トポロジーと、当該構造トポロジーに適合する生成文法とを対応付けて格納する文法格納ステップと、

RNA配列を上記生成文法に適用して構文解析木を導出する構文解析ステップと、

上記構文解析ステップにて導出された上記構文解析木に対して適合度の計算を行う適合度計算ステップと、

上記適合度計算ステップにより計算された上記適合度のうち予め定めた条件を満たす適合度である上記構文解析木を上記適合度が高い順にソートするソートステップと、

上記ソートステップによりソートされた上記構文解析木を上記RNA配列の二次構造の候補として出力する出力ステップと、

を含むことを特徴とするRNA配列解析方法。

【請求項8】 RNA二次構造の構造トポロジーと、当該構造トポロジーに適合する生成文法とを対応付けて格納する文法格納ステップと、

RNA配列を上記生成文法に適用して構文解析木を導出する構文解析ステップと、

上記構文解析ステップにて導出された上記構文解析木に対して適合度の計算を行う適合度計算ステップと、

上記適合度計算ステップにより計算された上記適合度のうち予め定めた条件を満たす適合度である上記構文解析木を導出した上記RNA配列を上記構造トポロジーの二次構造を持つRNA配列の候補として出力する出力ステップと、

を含むことを特徴とするRNA配列解析方法。

【請求項9】 RNA二次構造の構造トポロジーと、当該構造トポロジーに適合する生成文法とを対応付けて格納する文法格納ステップと、

RNA配列を上記生成文法に適用して構文解析木を導出する構文解析ステップと、

上記構文解析ステップにて導出された上記構文解析木に対して適合度の計算を行う適合度計算ステップと、

上記適合度計算ステップにより計算された上記適合度のうち予め定めた条件を満たす適合度である上記構文解析木を導出した上記RNA配列を抽出する抽出ステップと、

上記構造トポロジーと上記RNA配列とを二次元マトリックスで表示し、上記二次元マトリックスにおいて上記抽出ステップにて抽出された上記RNA配列と上記構造トポロジーに対応する格子部分をマークすることにより、上記RNA配列間で共通に有する構造トポロジーを可視化する共通構造マトリックス作成ステップと、

を含むことを特徴とするRNA配列解析方法。

【請求項10】 RNA二次構造の構造トポロジーと、当該構造トポロジーに適合する生成文法とを対応付けて格納する文法格納ステップと、

利用者が入力したDNA配列から転写されるRNA配列を作成するRNA配列

作成ステップと、

上記RNA配列作成ステップにより作成された上記RNA配列に対して上記生成文法を適用して構文解析木を導出する構文解析ステップと、

上記構文解析ステップにて導出された上記構文解析木に対して適合度の計算を行う適合度計算ステップと、

上記適合度計算ステップにより計算された上記適合度のうち予め定めた条件を満たす適合度である上記構文解析木を導出した上記RNA配列に対応する上記DNA配列部分を遺伝子の候補として予測する遺伝子予測ステップと、

を含むことを特徴とするRNA配列解析方法。

【請求項11】 RNA二次構造の構造トポロジーと、当該構造トポロジーに適合する生成文法とを対応付けて格納する文法格納ステップと、

RNA配列を上記生成文法に適用して構文解析木を導出する構文解析ステップと、

上記構文解析ステップにて導出された上記構文解析木に対して適合度の計算を行う適合度計算ステップと、

上記適合度計算ステップにより計算された上記適合度に基づいて上記RNA配列間の類似度を計算する類似度計算ステップと、

を含むことを特徴とするRNA配列解析方法。

【請求項12】 RNA二次構造の構造トポロジーと、当該構造トポロジーに適合する生成文法とを対応付けて格納する文法格納ステップと、

RNA配列を上記生成文法に適用して構文解析木を導出する構文解析ステップと、

上記構文解析ステップにて導出された上記構文解析木に対して適合度の計算を行う適合度計算ステップと、

上記適合度計算ステップにより計算された上記適合度のうち予め定めた条件を満たす適合度である上記構文解析木を導出した上記RNA配列を抽出する抽出ステップと、

上記構造トポロジーと上記RNA配列とを二次元マトリックスで表示し、上記二次元マトリックスにおいて上記抽出ステップにて抽出された上記RNA配列と

上記構造トポロジーに対応する格子部分に上記適合度を表示する適合度マトリックスを作成する適合度マトリックス作成ステップと、

上記適合度マトリックス作成ステップにて作成された上記適合度マトリックスについて、上記適合度により上記構造トポロジーをソートし、他のRNA配列について当該ソートされた上記構造トポロジーの順番に対応する上記生成文法により構文解析を行い上記適合度が最大となる上記構文解析木を求め、予め定めた条件を満たす上記適合度を持つ上記構文解析木に対応する上記他のRNA配列を抽出する共通構造抽出ステップと、

を含むことを特徴とするRNA配列解析方法。

【請求項13】 RNA二次構造の構造トポロジーと、当該構造トポロジーに適合する生成文法とを対応付けて格納する文法格納ステップと、

RNA配列を上記生成文法に適用して構文解析木を導出する構文解析ステップと、

上記構文解析ステップにて導出された上記構文解析木に対して適合度の計算を行う適合度計算ステップと、

上記適合度計算ステップにより計算された上記適合度のうち予め定めた条件を満たす適合度である上記構文解析木を上記適合度が高い順にソートするソートステップと、

上記ソートステップによりソートされた上記構文解析木を上記RNA配列の二次構造の候補として出力する出力ステップと、

を含むことを特徴とするRNA配列解析方法をコンピュータに実行させるプログラム。

【請求項14】 RNA二次構造の構造トポロジーと、当該構造トポロジーに適合する生成文法とを対応付けて格納する文法格納ステップと、

RNA配列を上記生成文法に適用して構文解析木を導出する構文解析ステップと、

上記構文解析ステップにて導出された上記構文解析木に対して適合度の計算を行う適合度計算ステップと、

上記適合度計算ステップにより計算された上記適合度のうち予め定めた条件を

満たす適合度である上記構文解析木を導出した上記RNA配列を上記構造トポロジーの二次構造を持つRNA配列の候補として出力する出力ステップと、

を含むことを特徴とするRNA配列解析方法をコンピュータに実行させるプログラム。

【請求項15】 RNA二次構造の構造トポロジーと、当該構造トポロジーに適合する生成文法とを対応付けて格納する文法格納ステップと、

RNA配列を上記生成文法に適用して構文解析木を導出する構文解析ステップと、

上記構文解析ステップにて導出された上記構文解析木に対して適合度の計算を行う適合度計算ステップと、

上記適合度計算ステップにより計算された上記適合度のうち予め定めた条件を満たす適合度である上記構文解析木を導出した上記RNA配列を抽出する抽出ステップと、

上記構造トポロジーと上記RNA配列とを二次元マトリックスで表示し、上記二次元マトリックスにおいて上記抽出ステップにて抽出された上記RNA配列と上記構造トポロジーに対応する格子部分をマークすることにより、上記RNA配列間で共通に有する構造トポロジーを可視化する共通構造マトリックス作成ステップと、

を含むことを特徴とするRNA配列解析方法をコンピュータに実行させるプログラム。

【請求項16】 RNA二次構造の構造トポロジーと、当該構造トポロジーに適合する生成文法とを対応付けて格納する文法格納ステップと、

利用者が入力したDNA配列から転写されるRNA配列を作成するRNA配列作成ステップと、

上記RNA配列作成ステップにより作成された上記RNA配列に対して上記生成文法を適用して構文解析木を導出する構文解析ステップと、

上記構文解析ステップにて導出された上記構文解析木に対して適合度の計算を行う適合度計算ステップと、

上記適合度計算ステップにより計算された上記適合度のうち予め定めた条件を

満たす適合度である上記構文解析木を導出した上記RNA配列に対応する上記DNA配列部分を遺伝子の候補として予測する遺伝子予測ステップと、

を含むことを特徴とするRNA配列解析方法をコンピュータに実行させるプログラム。

【請求項17】 RNA二次構造の構造トポロジーと、当該構造トポロジーに適合する生成文法とを対応付けて格納する文法格納ステップと、

RNA配列を上記生成文法に適用して構文解析木を導出する構文解析ステップと、

上記構文解析ステップにて導出された上記構文解析木に対して適合度の計算を行う適合度計算ステップと、

上記適合度計算ステップにより計算された上記適合度に基づいて上記RNA配列間の類似度を計算する類似度計算ステップと、

を含むことを特徴とするRNA配列解析方法をコンピュータに実行させるプログラム。

【請求項18】 RNA二次構造の構造トポロジーと、当該構造トポロジーに適合する生成文法とを対応付けて格納する文法格納ステップと、

RNA配列を上記生成文法に適用して構文解析木を導出する構文解析ステップと、

上記構文解析ステップにて導出された上記構文解析木に対して適合度の計算を行う適合度計算ステップと、

上記適合度計算ステップにより計算された上記適合度のうち予め定めた条件を満たす適合度である上記構文解析木を導出した上記RNA配列を抽出する抽出ステップと、

上記構造トポロジーと上記RNA配列とを二次元マトリックスで表示し、上記二次元マトリックスにおいて上記抽出ステップにて抽出された上記RNA配列と上記構造トポロジーに対応する格子部分に上記適合度を表示する適合度マトリックスを作成する適合度マトリックス作成ステップと、

上記適合度マトリックス作成ステップにて作成された上記適合度マトリックスについて、上記適合度により上記構造トポロジーをソートし、他のRNA配列に

について当該ソートされた上記構造トポロジーの順番に対応する上記生成文法により構文解析を行い上記適合度が最大となる上記構文解析木を求め、予め定めた条件を満たす上記適合度を持つ上記構文解析木に対応する上記他のRNA配列を抽出する共通構造抽出ステップと、

を含むことを特徴とするRNA配列解析方法をコンピュータに実行させるプログラム。

【請求項19】 上記請求項13～18のいずれか一つに記載されたプログラムを記録したことを特徴とするコンピュータ読み取り可能な記録媒体。

【発明の詳細な説明】

【0001】

【発明の属する技術分野】

本発明は、RNA配列解析装置、RNA配列解析方法、プログラム、および、記録媒体に関し、特に、RNAの二次構造を予測し、また、DNA配列から遺伝子部分を予測するRNA配列解析装置、RNA配列解析方法、プログラム、および、記録媒体に関する。

【0002】

【従来の技術】

RNA配列は、A（アデニン）、C（シトシン）、G（グアニン）、U（ウラシル）の4種の塩基により構成されるが、その一部は回文言語のような入れ子の状態となり、相補的な塩基同士（AとU、GとC、稀にGとU）が結合して二次構造を構成している。RNA配列の二次構造は、図1に示すように、多種の構造トポロジーを有している。ここで、積み重ねられた塩基対の連続領域をシステム（system）と呼び、また、塩基対で挟まれた一本鎖の部分配列をループ（loop）と呼ぶ。システムの端のループをヘアピンループという（図1のa）。システムの中にある一本鎖の塩基は、それがシステムの片側だけにあるときバルジループ（bulge loop）と呼び（図1のb）、システムの両側にあるときは内側ループ（internal loop）と呼ぶ（図1のc）。放射状に3個以上のシステムが出ているものをマルチループ（multi-branched loop）と呼ぶ。また、入れ子ではない塩基対があるとき、シュードノット（pseudo knot）

udeoknots) と呼ぶ(図1のd)。

【0003】

ここで、RNA配列を形式文法(生成文法)により構文解析することによりその二次構造を予測する手法が存在するが、正規文法では回文言語を記述することができないため、一般に、RNA二次構造解析においては、木文法(tree adjoining grammars)や、文脈自由文法(CFG)などを用いて構文解析を行い構造モデリング(構造トポロジー表現)を求める手法が考案されている。

【0004】

例えば、Yasuo Uemura等著「Tree adjoining grammars for RNA structure prediction (Theoretical Computer Science 210 1999 277p-303p)」(以下「文献1」という)には、木文法による構造モデリングと、パーザ(paser)を利用したエネルギー極小化によるRNA二次構造予測方法が開示されている。

【0005】

また、Elena Rivas and Sean R. Rddy著「The language of RNA: a formal grammar that includes pseudoknots (BIOINFORMATICS vol. 16 no. 4 2000 334p-340p)」(以下「文献2」という)には、Crossed-interaction Grammarsなどの独自の拡張を施した文脈自由文法(CFG)による構造モデリングと、パーザを利用したエネルギー極小化によるRNA二次構造予測方法が開示されている。

【0006】

また、Michael Zuker著「Prediction of RNA Secondary Structure by Energy minimization (July 8, 1996)」(以下「文献3」という)には、形式文法やパーザを用いた動的計画法(Dynamic Programming

g) によってRNA二次構造を予測する方法を用いたRNA配列解析システムであるMfold (製品名) が開示されている。これらの文献では、形式文法や動的計画法などの手法と、エネルギー極小化手法とを組み合わせることによってRNA二次構造予測精度を高めている。

【0007】

図2は、従来技術によるRNAの二次構造がステムループをとる場合の構文解析木の一例を示す図である。図2のaに示すRNA配列の二次構造を図2のbに、また、構文解析木を図2のcに示している。ここで、部分木 (subtree) は、内部の節を根とする構文解析木の断片である。RNA二次構造の構造トポロジーについて構文解析木を作成して構文解析を行うことにより二次構造解析を行う技術が研究されており、主要な構造トポロジーに対する文法が既知となっている。

【0008】

図3は、従来技術によるRNA二次構造の構造トポロジーについて、文法が固定されるとそれに対応した構造トポロジーが規定される（逆もまた可）ことを示す概念図である。ここで、生成文法（以下単に「文法」という）は、有限個の記号と、有限個の生成規則P (production rule) から成る。記号には、抽象的な非終端記号N (nonterminal symbol) と、観測文字列に実際に現れる終端記号T (terminal symbol) の2種類がある。終端記号Tは、RNA配列の場合にはA、T、G、Cの4文字である。図3に示すように、各構造トポロジーについてそれぞれ対応する文法を定義することができる。

【0009】

また、図4は、従来技術である木文法パーザを用いて、既知の文法からRNA配列の構文解析木を導出する場合の一例を示す図である。まず、構造未知のRNA配列を木文法パーザに入力する。ここで、木文法パーザは、入力された既知の木文法に従ってRNA配列の構文解析を行い構文解析木を導出する機能、および、導出された構文解析木について、ループや、塩基対とその他の二次構造要素のそれぞれの自由エネルギーの合計などを計算することにより平衡自由エネルギー

(ΔG) などの値を求める機能などを有する（文献1から3参照）。

【0010】

ここで、木文法パーザは必ずしも構文解析木を導出するわけではなく、入力したRNA配列が当該文法に適合しない場合（パーズが成功しない場合）には構文解析木を導出しない（すなわち、構文解析木は0個になる）。また、複数個の構文解析木が導出された場合には、エネルギー計算の結果、極小の自由エネルギーとなる1つの構文解析木を選択する。このとき、木文法パーザは導出過程の各段階において自由エネルギー極小な部分構造を見つけていくことができる。また、木文法パーザはエネルギー準最適な構文も出力することができる。このように、木文法パーザは、構文解析（パーズ）の途中でエネルギー計算を行うことにより、高速化と精度向上を実現している。

【0011】

【発明が解決しようとする課題】

しかしながら、従来の木文法パーザなどにより構文解析とエネルギー計算を行う手法を用いたRNA二次構造予測システムにおいては、RNA配列や抽出した文法を統合的に管理し、集積した文法やRNA配列を用いて二次構造予測などをより効率的に行うものは存在しなかったという問題点がある。

【0012】

また、与えられた特定の二次構造を取り得るようなRNA配列を検索するような方法は存在しなかったという問題点がある。

【0013】

また、複数のRNA配列に共通な二次構造を簡易に抽出する方法は存在しなかったという問題点がある。

【0014】

また、RNA配列からRNA二次構造に基づく類似度を簡易に求める方法は存在しなかったという問題点がある。

【0015】

さらに、DNA配列から遺伝子部分を発見するための手法としては、ホモロジー検索やモチーフ検索などを利用する手法が一般的であるが、未知の遺伝子部分

の発見には利用できないという問題点がある。ここで、従来技術で説明したように、RNA配列の構造トポロジーを予測可能な生成文法が求められているが、既知の生成文法により導出された構文解析木を用いた遺伝子発見方法は存在しなかったという問題点がある。

【0016】

このように、従来のシステム等は数々の問題点を有しており、その結果、システムの利用者および管理者のいずれにとっても、利便性が悪く、また、利用効率が悪いものであった。

本発明は上記問題点に鑑みてなされたもので、RNA配列や抽出した文法を統合的に管理し、集積した文法やRNA配列を用いて二次構造予測や新たな解析手法などをより効率的に行うことのできる、RNA配列解析装置、RNA配列解析方法、プログラム、および、記録媒体を提供することを目的としている。

【0017】

【課題を解決するための手段】

このような目的を達成するため、請求項1に記載のRNA配列解析装置は、RNA二次構造の構造トポロジーと、当該構造トポロジーに適合する生成文法とを対応付けて格納する文法格納手段と、RNA配列を上記生成文法に適用して構文解析木を導出する構文解析手段と、上記構文解析手段にて導出された上記構文解析木に対して適合度の計算を行う適合度計算手段と、上記適合度計算手段により計算された上記適合度のうち予め定めた条件を満たす適合度である上記構文解析木を上記適合度が高い順にソートするソート手段と、上記ソート手段によりソートされた上記構文解析木を上記RNA配列の二次構造の候補として出力する出力手段とを備えたことを特徴とする。

【0018】

この装置によれば、RNA二次構造の構造トポロジーと、当該構造トポロジーに適合する生成文法とを対応付けて格納し、RNA配列を生成文法に適用して構文解析木を導出し、導出された構文解析木に対して適合度の計算し、計算された適合度のうち予め定めた条件を満たす適合度である構文解析木を適合度が高い順にソートし、ソートされた構文解析木をRNA配列の二次構造の候補として出力

するので、一配列に対して多文法で構文解析を行うことができるようになる。すなわち、各生成文法に対してそれぞれ構文解析し適合度計算を行い適合度を得る。その結果、生成文法ごとに適合度が得られることになり、それらの適合度をソートすることによって生成文法に順位を付ける。これにより、生成文法に対する構造トポロジーにも順位が付けられることになるので、最終的にRNA配列が取り得る可能性の高い順に構造トポロジーを確認することができるようになる。

【0019】

また、請求項2に記載のRNA配列解析装置は、RNA二次構造の構造トポロジーと、当該構造トポロジーに適合する生成文法とを対応付けて格納する文法格納手段と、RNA配列を上記生成文法に適用して構文解析木を導出する構文解析手段と、上記構文解析手段にて導出された上記構文解析木に対して適合度の計算を行う適合度計算手段と、上記適合度計算手段により計算された上記適合度のうち予め定めた条件を満たす適合度である上記構文解析木を導出した上記RNA配列を上記構造トポロジーの二次構造を持つRNA配列の候補として出力する出力手段とを備えたことを特徴とする。

【0020】

この装置によれば、RNA二次構造の構造トポロジーと、当該構造トポロジーに適合する生成文法とを対応付けて格納し、RNA配列を生成文法に適用して構文解析木を導出し、導出された構文解析木に対して適合度の計算を行い、計算された適合度のうち予め定めた条件を満たす適合度である構文解析木を導出したRNA配列を上記構造トポロジーの二次構造を持つRNA配列の候補として出力するので、多配列に対して一文法で構文解析を行うことができるようになる。すなわち、与えられた特定の構造トポロジーに対し、対応する生成文法を取得し、これを用いてRNA配列データベースに格納されているすべてまたは一部のRNA配列をそれぞれ構文解析し、ある閾値以下の適合度で構文解析に成功したRNA配列群を結果として出力する。これにより、与えられた特定の二次構造を取り得るようなRNA配列を検索することができるようになる。

【0021】

また、請求項3に記載のRNA配列解析装置は、RNA二次構造の構造トポロ

ジーと、当該構造トポロジーに適合する生成文法とを対応付けて格納する文法格納手段と、RNA配列を上記生成文法に適用して構文解析木を導出する構文解析手段と、上記構文解析手段にて導出された上記構文解析木に対して適合度の計算を行う適合度計算手段と、上記適合度計算手段により計算された上記適合度のうち予め定めた条件を満たす適合度である上記構文解析木を導出した上記RNA配列を抽出する抽出手段と、上記構造トポロジーと上記RNA配列とを二次元マトリックスで表示し、上記二次元マトリックスにおいて上記抽出手段にて抽出された上記RNA配列と上記構造トポロジーに対応する格子部分をマークすることにより、上記RNA配列間で共通に有する構造トポロジーを可視化する共通構造マトリックス作成手段とを備えたことを特徴とする。

【0022】

この装置によれば、RNA二次構造の構造トポロジーと、当該構造トポロジーに適合する生成文法とを対応付けて格納し、RNA配列を生成文法に適用して構文解析木を導出し、導出された構文解析木に対して適合度の計算を行い、計算された適合度のうち予め定めた条件を満たす適合度である構文解析木を導出したRNA配列を抽出し、構造トポロジーとRNA配列とを二次元マトリックスで表示し、二次元マトリックスにおいて抽出されたRNA配列と構造トポロジーに対応する格子部分をマークすることにより、RNA配列間で共通に有する構造トポロジーを可視化するので、RNA配列間の共通構造を容易に発見することができるようになる。

【0023】

また、請求項4に記載のRNA配列解析装置は、RNA二次構造の構造トポロジーと、当該構造トポロジーに適合する生成文法とを対応付けて格納する文法格納手段と、利用者が入力したDNA配列から転写されるRNA配列を作成するRNA配列作成手段と、上記RNA配列作成手段により作成された上記RNA配列に対して上記生成文法を適用して構文解析木を導出する構文解析手段と、上記構文解析手段にて導出された上記構文解析木に対して適合度の計算を行う適合度計算手段と、上記適合度計算手段により計算された上記適合度のうち予め定めた条件を満たす適合度である上記構文解析木を導出した上記RNA配列に対応する上

記DNA配列部分を遺伝子の候補として予測する遺伝子予測手段とを備えたことを特徴とする。

【0024】

この装置によれば、RNA二次構造の構造トポロジーと、当該構造トポロジーに適合する生成文法とを対応付けて格納し、利用者が入力したDNA配列から転写されるRNA配列を作成し、作成されたRNA配列に対して生成文法を適用して構文解析木を導出し、導出された構文解析木に対して適合度の計算を行い、計算された適合度のうち予め定めた条件を満たす適合度である構文解析木を導出したRNA配列に対応するDNA配列部分を遺伝子の候補として予測するので、DNA配列のうち既知のトポロジーを有する可能性のあるRNA配列に対応する部分について、遺伝子部分である可能性があることを予測することができるようになる。

【0025】

また、請求項5に記載のRNA配列解析装置は、RNA二次構造の構造トポロジーと、当該構造トポロジーに適合する生成文法とを対応付けて格納する文法格納手段と、RNA配列を上記生成文法に適用して構文解析木を導出する構文解析手段と、上記構文解析手段にて導出された上記構文解析木に対して適合度の計算を行う適合度計算手段と、上記適合度計算手段により計算された上記適合度に基づいて上記RNA配列間の類似度を計算する類似度計算手段とを備えたことを特徴とする。

【0026】

この装置によれば、RNA二次構造の構造トポロジーと、当該構造トポロジーに適合する生成文法とを対応付けて格納し、RNA配列を生成文法に適用して構文解析木を導出し、導出された構文解析木に対して適合度の計算を行い、計算された適合度に基づいてRNA配列間の類似度を計算するので、RNA構造の類似度を容易に求めることができるようになる。

【0027】

また、請求項6に記載のRNA配列解析装置は、RNA二次構造の構造トポロジーと、当該構造トポロジーに適合する生成文法とを対応付けて格納する文法格

納手段と、RNA配列を上記生成文法に適用して構文解析木を導出する構文解析手段と、上記構文解析手段にて導出された上記構文解析木に対して適合度の計算を行う適合度計算手段と、上記適合度計算手段により計算された上記適合度のうち予め定めた条件を満たす適合度である上記構文解析木を導出した上記RNA配列を抽出する抽出手段と、上記構造トポロジーと上記RNA配列とを二次元マトリックスで表示し、上記二次元マトリックスにおいて上記抽出手段にて抽出された上記RNA配列と上記構造トポロジーに対応する格子部分に上記適合度を表示する適合度マトリックスを作成する適合度マトリックス作成手段と、上記適合度マトリックス作成手段にて作成された上記適合度マトリックスについて、上記適合度により上記構造トポロジーをソートし、他のRNA配列について当該ソートされた上記構造トポロジーの順番に対応する上記生成文法により構文解析を行い上記適合度が最大となる上記構文解析木を求め、予め定めた条件を満たす上記適合度を持つ上記構文解析木に対応する上記他のRNA配列を抽出する共通構造抽出手段とを備えたことを特徴とする。

【0028】

この装置によれば、RNA二次構造の構造トポロジーと、当該構造トポロジーに適合する生成文法とを対応付けて格納し、RNA配列を生成文法に適用して構文解析木を導出し、導出された構文解析木に対して適合度の計算を行い、計算された適合度のうち予め定めた条件を満たす適合度である構文解析木を導出したRNA配列を抽出し、構造トポロジーとRNA配列とを三次元マトリックスで表示し、二次元マトリックスにおいて抽出されたRNA配列と構造トポロジーに対応する格子部分に適合度を表示する適合度マトリックスを作成し、適合度マトリックスについて、適合度により構造トポロジーをソートし、他のRNA配列について当該ソートされた構造トポロジーの順番に対応する生成文法により構文解析を行い適合度が最大となる構文解析木を求め、予め定めた条件を満たす適合度を持つ構文解析木に対応する他のRNA配列を抽出するので、共通の構造を持つRNA配列を容易に発見することができるようになる。

【0029】

また、本発明はRNA配列解析方法に関するものであり、請求項7に記載のR

RNA配列解析方法は、RNA二次構造の構造トポロジーと、当該構造トポロジーに適合する生成文法とを対応付けて格納する文法格納ステップと、RNA配列を上記生成文法に適用して構文解析木を導出する構文解析ステップと、上記構文解析ステップにて導出された上記構文解析木に対して適合度の計算を行う適合度計算ステップと、上記適合度計算ステップにより計算された上記適合度のうち予め定めた条件を満たす適合度である上記構文解析木を上記適合度が高い順にソートするソートステップと、上記ソートステップによりソートされた上記構文解析木を上記RNA配列の二次構造の候補として出力する出力ステップとを含むことを特徴とする。

【0030】

この方法によれば、RNA二次構造の構造トポロジーと、当該構造トポロジーに適合する生成文法とを対応付けて格納し、RNA配列を生成文法に適用して構文解析木を導出し、導出された構文解析木に対して適合度を計算し、計算された適合度のうち予め定めた条件を満たす適合度である構文解析木を適合度が高い順にソートし、ソートされた構文解析木をRNA配列の二次構造の候補として出力するので、一配列に対して多文法で構文解析を行うことができるようになる。すなわち、各生成文法に対してそれぞれ構文解析し適合度計算を行い適合度を得る。その結果、生成文法ごとに適合度が得られることになり、それらの適合度をソートすることによって生成文法に順位を付ける。これにより、生成文法に対する構造トポロジーにも順位が付けられることになるので、最終的にRNA配列が取り得る可能性の高い順に構造トポロジーを確認することができるようになる。

【0031】

また、請求項8に記載のRNA配列解析方法は、RNA二次構造の構造トポロジーと、当該構造トポロジーに適合する生成文法とを対応付けて格納する文法格納ステップと、RNA配列を上記生成文法に適用して構文解析木を導出する構文解析ステップと、上記構文解析ステップにて導出された上記構文解析木に対して適合度の計算を行う適合度計算ステップと、上記適合度計算ステップにより計算された上記適合度のうち予め定めた条件を満たす適合度である上記構文解析木を導出した上記RNA配列を上記構造トポロジーの二次構造を持つRNA配列の候

補として出力する出力ステップとを含むことを特徴とする。

【0032】

この方法によれば、RNA二次構造の構造トポロジーと、当該構造トポロジーに適合する生成文法とを対応付けて格納し、RNA配列を生成文法に適用して構文解析木を導出し、導出された構文解析木に対して適合度の計算を行い、計算された適合度のうち予め定めた条件を満たす適合度である構文解析木を導出したRNA配列を上記構造トポロジーの二次構造を持つRNA配列の候補として出力するので、多配列に対して一文法で構文解析を行うことができるようになる。すなわち、与えられた特定の構造トポロジーに対し、対応する生成文法を取得し、これを用いてRNA配列データベースに格納されているすべてまたは一部のRNA配列をそれぞれ構文解析し、ある閾値以下の適合度で構文解析に成功したRNA配列群を結果として出力する。これにより、与えられた特定の二次構造を取り得るようなRNA配列を検索することができるようになる。

【0033】

また、請求項9に記載のRNA配列解析方法は、RNA二次構造の構造トポロジーと、当該構造トポロジーに適合する生成文法とを対応付けて格納する文法格納ステップと、RNA配列を上記生成文法に適用して構文解析木を導出する構文解析ステップと、上記構文解析ステップにて導出された上記構文解析木に対して適合度の計算を行う適合度計算ステップと、上記適合度計算ステップにより計算された上記適合度のうち予め定めた条件を満たす適合度である上記構文解析木を導出した上記RNA配列を抽出する抽出ステップと、上記構造トポロジーと上記RNA配列とを二次元マトリックスで表示し、上記二次元マトリックスにおいて上記抽出ステップにて抽出された上記RNA配列と上記構造トポロジーに対応する格子部分をマークすることにより、上記RNA配列間で共通に有する構造トポロジーを可視化する共通構造マトリックス作成ステップとを含むことを特徴とする。

【0034】

この方法によれば、RNA二次構造の構造トポロジーと、当該構造トポロジーに適合する生成文法とを対応付けて格納し、RNA配列を生成文法に適用して構

文解析木を導出し、導出された構文解析木に対して適合度の計算を行い、計算された適合度のうち予め定めた条件を満たす適合度である構文解析木を導出したRNA配列を抽出し、構造トポロジーとRNA配列とを二次元マトリックスで表示し、二次元マトリックスにおいて抽出されたRNA配列と構造トポロジーに対応する格子部分をマークすることにより、RNA配列間で共通に有する構造トポロジーを可視化するので、RNA配列間の共通構造を容易に発見することができるようになる。

【0035】

また、請求項10に記載のRNA配列解析方法は、RNA二次構造の構造トポロジーと、当該構造トポロジーに適合する生成文法とを対応付けて格納する文法格納ステップと、利用者が入力したDNA配列から転写されるRNA配列を作成するRNA配列作成ステップと、上記RNA配列作成ステップにより作成された上記RNA配列に対して上記生成文法を適用して構文解析木を導出する構文解析ステップと、上記構文解析ステップにて導出された上記構文解析木に対して適合度の計算を行う適合度計算ステップと、上記適合度計算ステップにより計算された上記適合度のうち予め定めた条件を満たす適合度である上記構文解析木を導出した上記RNA配列に対応する上記DNA配列部分を遺伝子の候補として予測する遺伝子予測ステップとを含むことを特徴とする。

【0036】

この方法によれば、RNA二次構造の構造トポロジーと、当該構造トポロジーに適合する生成文法とを対応付けて格納し、利用者が入力したDNA配列から転写されるRNA配列を作成し、作成されたRNA配列に対して生成文法を適用して構文解析木を導出し、導出された構文解析木に対して適合度の計算を行い、計算された適合度のうち予め定めた条件を満たす適合度である構文解析木を導出したRNA配列に対応するDNA配列部分を遺伝子の候補として予測するので、DNA配列のうち既知のトポロジーを有する可能性のあるRNA配列に対応する部分について、遺伝子部分である可能性があることを予測することができるようになる。

【0037】

また、請求項11に記載のRNA配列解析方法は、RNA二次構造の構造トポロジーと、当該構造トポロジーに適合する生成文法とを対応付けて格納する文法格納ステップと、RNA配列を上記生成文法に適用して構文解析木を導出する構文解析ステップと、上記構文解析ステップにて導出された上記構文解析木に対して適合度の計算を行う適合度計算ステップと、上記適合度計算ステップにより計算された上記適合度に基づいて上記RNA配列間の類似度を計算する類似度計算ステップとを含むことを特徴とする。

【0038】

この方法によれば、RNA二次構造の構造トポロジーと、当該構造トポロジーに適合する生成文法とを対応付けて格納し、RNA配列を生成文法に適用して構文解析木を導出し、導出された構文解析木に対して適合度の計算を行い、計算された適合度に基づいてRNA配列間の類似度を計算するので、RNA構造の類似度を容易に求めることができるようになる。

【0039】

また、請求項12に記載のRNA配列解析方法は、RNA二次構造の構造トポロジーと、当該構造トポロジーに適合する生成文法とを対応付けて格納する文法格納ステップと、RNA配列を上記生成文法に適用して構文解析木を導出する構文解析ステップと、上記構文解析ステップにて導出された上記構文解析木に対して適合度の計算を行う適合度計算ステップと、上記適合度計算ステップにより計算された上記適合度のうち予め定めた条件を満たす適合度である上記構文解析木を導出した上記RNA配列を抽出する抽出ステップと、上記構造トポロジーと上記RNA配列とを二次元マトリックスで表示し、上記二次元マトリックスにおいて上記抽出ステップにて抽出された上記RNA配列と上記構造トポロジーに対応する格子部分に上記適合度を表示する適合度マトリックスを作成する適合度マトリックス作成ステップと、上記適合度マトリックス作成ステップにて作成された上記適合度マトリックスについて、上記適合度により上記構造トポロジーをソートし、他のRNA配列について当該ソートされた上記構造トポロジーの順番に対応する上記生成文法により構文解析を行い上記適合度が最大となる上記構文解析木を求め、予め定めた条件を満たす上記適合度を持つ上記構文解析木に対応する

上記他のRNA配列を抽出する共通構造抽出ステップとを含むことを特徴とする。

【0040】

この方法によれば、RNA二次構造の構造トポロジーと、当該構造トポロジーに適合する生成文法とを対応付けて格納し、RNA配列を生成文法に適用して構文解析木を導出し、導出された構文解析木に対して適合度の計算を行い、計算された適合度のうち予め定めた条件を満たす適合度である構文解析木を導出したRNA配列を抽出し、構造トポロジーとRNA配列とを二次元マトリックスで表示し、二次元マトリックスにおいて抽出されたRNA配列と構造トポロジーに対応する格子部分に適合度を表示する適合度マトリックスを作成し、適合度マトリックスについて、適合度により構造トポロジーをソートし、他のRNA配列について当該ソートされた構造トポロジーの順番に対応する生成文法により構文解析を行い適合度が最大となる構文解析木を求め、予め定めた条件を満たす適合度を持つ構文解析木に対応する他のRNA配列を抽出するので、共通の構造を持つRNA配列を容易に発見することができるようになる。

【0041】

また、本発明はRNA配列解析方法をコンピュータに実行させるプログラムに関するものであり、請求項13に記載のプログラムは、RNA二次構造の構造トポロジーと、当該構造トポロジーに適合する生成文法とを対応付けて格納する文法格納ステップと、RNA配列を上記生成文法に適用して構文解析木を導出する構文解析ステップと、上記構文解析ステップにて導出された上記構文解析木に対して適合度の計算を行う適合度計算ステップと、上記適合度計算ステップにより計算された上記適合度のうち予め定めた条件を満たす適合度である上記構文解析木を上記適合度が高い順にソートするソートステップと、上記ソートステップによりソートされた上記構文解析木を上記RNA配列の二次構造の候補として出力する出力ステップとを含むことを特徴とする。

【0042】

このプログラムによれば、RNA二次構造の構造トポロジーと、当該構造トポロジーに適合する生成文法とを対応付けて格納し、RNA配列を生成文法に適用

して構文解析木を導出し、導出された構文解析木に対して適合度の計算し、計算された適合度のうち予め定めた条件を満たす適合度である構文解析木を適合度が高い順にソートし、ソートされた構文解析木をRNA配列の二次構造の候補として出力するので、一配列に対して多文法で構文解析を行うことができるようになる。すなわち、各生成文法に対してそれぞれ構文解析し適合度計算を行い適合度を得る。その結果、生成文法ごとに適合度が得られることになり、それらの適合度をソートすることによって生成文法に順位を付ける。これにより、生成文法に対する構造トポロジーにも順位が付けられることになるので、最終的にRNA配列が取り得る可能性の高い順に構造トポロジーを確認することができるようになる。

【0043】

また、請求項14に記載のプログラムは、RNA二次構造の構造トポロジーと、当該構造トポロジーに適合する生成文法とを対応付けて格納する文法格納ステップと、RNA配列を上記生成文法に適用して構文解析木を導出する構文解析ステップと、上記構文解析ステップにて導出された上記構文解析木に対して適合度の計算を行う適合度計算ステップと、上記適合度計算ステップにより計算された上記適合度のうち予め定めた条件を満たす適合度である上記構文解析木を導出した上記RNA配列を上記構造トポロジーの二次構造を持つRNA配列の候補として出力する出力ステップとを含むことを特徴とする。

【0044】

このプログラムによれば、RNA二次構造の構造トポロジーと、当該構造トポロジーに適合する生成文法とを対応付けて格納し、RNA配列を生成文法に適用して構文解析木を導出し、導出された構文解析木に対して適合度の計算を行い、計算された適合度のうち予め定めた条件を満たす適合度である構文解析木を導出したRNA配列を上記構造トポロジーの二次構造を持つRNA配列の候補として出力するので、多配列に対して一文法で構文解析を行うことができるようになる。すなわち、与えられた特定の構造トポロジーに対し、対応する生成文法を取得し、これを用いてRNA配列データベースに格納されているすべてまたは一部のRNA配列をそれぞれ構文解析し、ある閾値以下の適合度で構文解析に成功した

RNA配列群を結果として出力する。これにより、与えられた特定の二次構造を取り得るようなRNA配列を検索することができるようになる。

【0045】

また、請求項15に記載のプログラムは、RNA二次構造の構造トポロジーと、当該構造トポロジーに適合する生成文法とを対応付けて格納する文法格納ステップと、RNA配列を上記生成文法に適用して構文解析木を導出する構文解析ステップと、上記構文解析ステップにて導出された上記構文解析木に対して適合度の計算を行う適合度計算ステップと、上記適合度計算ステップにより計算された上記適合度のうち予め定めた条件を満たす適合度である上記構文解析木を導出した上記RNA配列を抽出する抽出ステップと、上記構造トポロジーと上記RNA配列とを二次元マトリックスで表示し、上記二次元マトリックスにおいて上記抽出ステップにて抽出された上記RNA配列と上記構造トポロジーに対応する格子部分をマークすることにより、上記RNA配列間で共通に有する構造トポロジーを可視化する共通構造マトリックス作成ステップとを含むことを特徴とする。

【0046】

このプログラムによれば、RNA二次構造の構造トポロジーと、当該構造トポロジーに適合する生成文法とを対応付けて格納し、RNA配列を生成文法に適用して構文解析木を導出し、導出された構文解析木に対して適合度の計算を行い、計算された適合度のうち予め定めた条件を満たす適合度である構文解析木を導出したRNA配列を抽出し、構造トポロジーとRNA配列とを二次元マトリックスで表示し、二次元マトリックスにおいて抽出されたRNA配列と構造トポロジーに対応する格子部分をマークすることにより、RNA配列間で共通に有する構造トポロジーを可視化するので、RNA配列間の共通構造を容易に発見することができるようになる。

【0047】

また、請求項16に記載のプログラムは、RNA二次構造の構造トポロジーと、当該構造トポロジーに適合する生成文法とを対応付けて格納する文法格納ステップと、利用者が入力したDNA配列から転写されるRNA配列を作成するRNA配列作成ステップと、上記RNA配列作成ステップにより作成された上記RN

A配列に対して上記生成文法を適用して構文解析木を導出する構文解析ステップと、上記構文解析ステップにて導出された上記構文解析木に対して適合度の計算を行う適合度計算ステップと、上記適合度計算ステップにより計算された上記適合度のうち予め定めた条件を満たす適合度である上記構文解析木を導出した上記RNA配列に対応する上記DNA配列部分を遺伝子の候補として予測する遺伝子予測ステップとを含むことを特徴とする。

【0048】

このプログラムによれば、RNA二次構造の構造トポロジーと、当該構造トポロジーに適合する生成文法とを対応付けて格納し、利用者が入力したDNA配列から転写されるRNA配列を作成し、作成されたRNA配列に対して生成文法を適用して構文解析木を導出し、導出された構文解析木に対して適合度の計算を行い、計算された適合度のうち予め定めた条件を満たす適合度である構文解析木を導出したRNA配列に対応するDNA配列部分を遺伝子の候補として予測するので、DNA配列のうち既知のトポロジーを有する可能性のあるRNA配列に対応する部分について、遺伝子部分である可能性があることを予測することができるようになる。

【0049】

また、請求項17に記載のプログラムは、RNA二次構造の構造トポロジーと、当該構造トポロジーに適合する生成文法とを対応付けて格納する文法格納ステップと、RNA配列を上記生成文法に適用して構文解析木を導出する構文解析ステップと、上記構文解析ステップにて導出された上記構文解析木に対して適合度の計算を行う適合度計算ステップと、上記適合度計算ステップにより計算された上記適合度に基づいて上記RNA配列間の類似度を計算する類似度計算ステップとを含むことを特徴とする。

【0050】

このプログラムによれば、RNA二次構造の構造トポロジーと、当該構造トポロジーに適合する生成文法とを対応付けて格納し、RNA配列を生成文法に適用して構文解析木を導出し、導出された構文解析木に対して適合度の計算を行い、計算された適合度に基づいてRNA配列間の類似度を計算するので、RNA構造

の類似度を容易に求めることができるようになる。

【005.1】

また、請求項18に記載のプログラムは、RNA二次構造の構造トポロジーと、当該構造トポロジーに適合する生成文法とを対応付けて格納する文法格納ステップと、RNA配列を上記生成文法に適用して構文解析木を導出する構文解析ステップと、上記構文解析ステップにて導出された上記構文解析木に対して適合度の計算を行う適合度計算ステップと、上記適合度計算ステップにより計算された上記適合度のうち予め定めた条件を満たす適合度である上記構文解析木を導出した上記RNA配列を抽出する抽出ステップと、上記構造トポロジーと上記RNA配列とを二次元マトリックスで表示し、上記二次元マトリックスにおいて上記抽出ステップにて抽出された上記RNA配列と上記構造トポロジーに対応する格子部分に上記適合度を表示する適合度マトリックスを作成する適合度マトリックス作成ステップと、上記適合度マトリックス作成ステップにて作成された上記適合度マトリックスについて、上記適合度により上記構造トポロジーをソートし、他のRNA配列について当該ソートされた上記構造トポロジーの順番に対応する上記生成文法により構文解析を行い上記適合度が最大となる上記構文解析木を求め、予め定めた条件を満たす上記適合度を持つ上記構文解析木に対応する上記他のRNA配列を抽出する共通構造抽出ステップとを含むことを特徴とする。

【005.2】

このプログラムによれば、RNA二次構造の構造トポロジーと、当該構造トポロジーに適合する生成文法とを対応付けて格納し、RNA配列を生成文法に適用して構文解析木を導出し、導出された構文解析木に対して適合度の計算を行い、計算された適合度のうち予め定めた条件を満たす適合度である構文解析木を導出したRNA配列を抽出し、構造トポロジーとRNA配列とを二次元マトリックスで表示し、二次元マトリックスにおいて抽出されたRNA配列と構造トポロジーに対応する格子部分に適合度を表示する適合度マトリックスを作成し、適合度マトリックスについて、適合度により構造トポロジーをソートし、他のRNA配列について当該ソートされた構造トポロジーの順番に対応する生成文法により構文解析を行い適合度が最大となる構文解析木を求め、予め定めた条件を満たす適合

度を持つ構文解析木に対応する他のRNA配列を抽出するので、共通の構造を持つRNA配列を容易に発見することができるようになる。

【0053】

また、本発明は記録媒体に関するものであり、請求項19に記載の記録媒体は、上記請求項13から18のいずれか一つに記載されたプログラムを記録したことを特徴とする。

【0054】

この記録媒体によれば、当該記録媒体に記録されたプログラムをコンピュータに読み取らせて実行することによって、請求項13から18のいずれか一つに記載されたプログラムをコンピュータを利用して実現することができ、これら各プログラムと同様の効果を得ることができる。

【0055】

【発明の実施の形態】

以下に、本発明にかかるRNA配列解析装置、RNA配列解析方法、プログラム、および、記録媒体の実施の形態を図面に基づいて詳細に説明する。なお、この実施の形態によりこの発明が限定されるものではない。

特に、以下の実施の形態においては、本発明を、木文法に適用した例について説明するが、この場合に限らず、全ての生成文法において、同様に適用することができる。

【0056】

【本システムの概要】

以下、本システムの概要について説明し、その後、本システムの構成および処理等について詳細に説明する。

このシステムは、概略的に、以下の基本的特徴を有する。すなわち、本システムのRNA配列解析装置は、RNA二次構造の構造トポロジーと、当該構造トポロジーに適合する生成文法とを対応付けて格納し、RNA配列を生成文法に適用して構文解析木を導出し、導出された構文解析木に対して適合度を計算し、計算された適合度のうち予め定めた条件を満たす適合度である構文解析木を適合度が高い順にソートし、ソートされた構文解析木をRNA配列の二次構造の候補とし

て出力する。ここで、生成文法は、木文法、文脈自由文法などを含むが、シードノットを表現するためには木文法が最も適しているため、木文法を用いることが好ましい。

【0057】

また、本装置は、導出された構文解析木に対して適合度の計算を行い、計算された適合度のうち予め定めた条件を満たす適合度である構文解析木を導出したRNA配列を上記構造トポロジーの二次構造を持つRNA配列の候補として出力する。

【0058】

また、本装置は、計算された適合度のうち予め定めた条件を満たす適合度である構文解析木を導出したRNA配列を抽出し、構造トポロジーとRNA配列とを二次元マトリックスで表示し、二次元マトリックスにおいて抽出されたRNA配列と構造トポロジーに対応する格子部分をマークすることにより、RNA配列間で共通に有する構造トポロジーを可視化する。

【0059】

また、本装置は、利用者が入力したDNA配列から転写されるRNA配列を作成し、作成されたRNA配列に対して生成文法を適用して構文解析木を導出し、導出された構文解析木に対して適合度の計算を行い、計算された適合度のうち予め定めた条件を満たす適合度である構文解析木を導出したRNA配列に対応するDNA配列部分を遺伝子の候補として予測する。

【0060】

さらに、本装置は、RNA二次構造の構造トポロジーと、当該構造トポロジーに適合する生成文法とを対応付けて格納し、RNA配列を生成文法に適用して構文解析木を導出し、導出された構文解析木に対して適合度の計算を行い、計算された適合度に基づいてRNA配列間の類似度を計算する。

【0061】

【システム構成】

まず、本システムの構成について説明する。図5は、本発明が適用される本システムの構成の一例を示すブロック図であり、該構成のうち本発明に関係する部

分のみを概念的に示している。本システムは、概略的に、配列情報を解析するRNA配列解析装置であるRNA配列解析装置100と、配列情報等に関する外部データベースやホモロジー検索用の外部分析プログラム等を提供する外部システム200とを、ネットワーク300を介して通信可能に接続して構成されている。

【0062】

図5においてネットワーク300は、RNA配列解析装置100と外部システム200とを相互に接続する機能を有し、例えば、インターネット等である。

【0063】

図5において外部システム200は、ネットワーク300を介して、RNA配列解析装置100と相互に接続され、利用者に対して配列情報等に関する外部データベースやホモロジー検索やモチーフ検索等の外部分析プログラムを実行するウェブサイトを提供する機能を有する。

【0064】

ここで、外部システム200は、WEBサーバやASPサーバ等として構成してもよく、そのハードウェア構成は、一般に市販されるワークステーション、パソコン用コンピュータ等の情報処理装置およびその付属装置により構成してもよい。また、外部システム200の各機能は、外部システム200のハードウェア構成中のCPU、ディスク装置、メモリ装置、入力装置、出力装置、通信制御装置等およびそれらを制御するプログラム等により実現される。

【0065】

図5においてRNA配列解析装置100は、概略的に、RNA配列解析装置100の全体を統括的に制御するCPU等の制御部102、通信回線等に接続されるルータ等の通信装置（図示せず）に接続される通信制御インターフェース部104、入力装置112および出力装置114に接続される入出力制御インターフェース部108、および、各種のデータベースやテーブル（RNA配列データベース106a～共通構造マトリックス106c）を格納する記憶部106を備えて構成されており、これら各部は任意の通信路を介して通信可能に接続されている。さらに、このRNA配列解析装置100は、ルータ等の通信装置および専用

線等の有線または無線の通信回線を介して、ネットワーク300に通信可能に接続されている。

【0066】

記憶部106に格納される各種のデータベース（RNA配列データベース106a～共通構造マトリックス106c）は、固定ディスク装置等のストレージ手段であり、各種処理に用いる各種のプログラムやテーブルやファイルやデータベースやウェブページ用ファイル等を格納する。

【0067】

これら記憶部106の各構成要素のうち、RNA配列データベース106aは、RNA配列を格納したデータベースである。RNA配列データベース106aは、インターネットを経由してアクセスする外部のRNA配列データベースであってもよく、また、これらのデータベースをコピーしたり、オリジナルの配列情報を格納したり、さらに独自のアノテーション情報等を付加したりして作成したインハウスデータベースであってもよい。また、RNA配列データベース106aは、cDNA等のDNA配列データベースに基づいて予め生成された、あるいは必要時に動的に生成されたRNA配列を格納したものでもよい。

【0068】

また、文法データベース106bは、RNA二次構造の構造トポロジーと、当該構造トポロジーに適合する生成文法とを対応付けて格納する文法格納手段である。ここで、図6は、文法データベース106bに格納される情報の一例を示す図である。図6に示すように、文法データベース106bは、構造トポロジーと、その構造トポロジーに対応する文法とを対応付けて格納する。ここで、文法データベース106bには、図6に示したように、構造トポロジーと文法とが1対1で対応するようにしてもよく、また、複数のトポロジーが結合した文法（例えば、シードノットとヘアピンループとを両方持つトポロジーなど）や、特徴的な構造を有するRNA用の文法（例えば、rRNAに特有の構造トポロジーなど）や、所定のカテゴリーのRNAが共通で備えるトポロジー用の文法や、全てのRNAに適合する文法などを規定してもよい。

【0069】

また、共通構造マトリックス106cは、構造トポロジーとRNA配列とを二次元マトリックスで表示するためのテーブル（記憶領域）である。

【0070】

また、図5において、通信制御インターフェース部104は、RNA配列解析装置100とネットワーク300（またはルータ等の通信装置）との間における通信制御を行う。すなわち、通信制御インターフェース部104は、他の端末と通信回線を介してデータを通信する機能を有する。

【0071】

また、図5において、入出力制御インターフェース部108は、入力装置112や出力装置114の制御を行う。ここで、出力装置114としては、モニタ（家庭用テレビを含む）の他、スピーカを用いることができる（なお、以下においては出力装置をモニタとして記載する）。また、入力装置112としては、キーボード、マウス、および、マイク等を用いることができる。また、モニタも、マウスと協働してポインティングデバイス機能を実現する。

【0072】

また、図5において、制御部102は、OS(Operating System)等の制御プログラム、各種の処理手順等を規定したプログラム、および所要データを格納するための内部メモリを有し、これらのプログラム等により、種々の処理を実行するための情報処理を行う。制御部102は、機能概念的に、構造予測部102a、類似度計算部102d、共通構造マトリックス作成部102f、および、遺伝子予測部102gを備えて構成されている。

【0073】

このうち、構造予測部102aは、入力された既知の文法に従ってRNA配列の構文解析を行い構文解析木を導出する機能（構文解析部102b）、および、導出された構文解析木に対して適合度の計算を行う機能（適合度計算部102c）などを有する。

【0074】

また、類似度計算部102dは、複数のRNA配列間の類似度を計算する類似度計算手段である。

【0075】

また、共通構造マトリックス作成部102fは、適合度計算手段により計算された適合度のうち予め定めた条件を満たす適合度である構文解析木を導出したRNA配列を抽出する抽出手段、および、構造トポロジーとRNA配列とを二次元マトリックスで表示し、二次元マトリックスにおいて抽出手段にて抽出されたRNA配列と構造トポロジーに対応する格子部分をマークすることにより、RNA配列間で共通に有する構造トポロジーを可視化する共通構造マトリックス作成手段、二次元マトリックスにおいて抽出手段にて抽出されたRNA配列と構造トポロジーに対応する格子部分に適合度を表示する適合度マトリックスを作成する適合度マトリックス作成手段、および、適合度マトリックス作成手段にて作成された適合度マトリックスについて、適合度により構造トポロジーをソートし、他のRNA配列について当該ソートされた構造トポロジーの順番に対応する生成文法により構文解析を行い適合度が最大となる構文解析木を求め、予め定めた条件を満たす適合度を持つ構文解析木に対応する他のRNA配列を抽出する共通構造抽出手段である。

【0076】

また、遺伝子予測部102gは、利用者が入力したDNA配列から転写されるRNA配列を作成するRNA配列作成手段、および、適合度計算手段により計算された適合度のうち予め定めた条件を満たす適合度である構文解析木を導出したRNA配列に対応するDNA配列部分を遺伝子の候補として予測する遺伝子予測手段である。なお、これら各部によって行なわれる処理の詳細については、後述する。

【0077】

[システムの処理]

次に、このように構成された本実施の形態における本システムの処理の一例について、以下に図7～図11を参照して詳細に説明する。

【0078】

[RNA二次構造予測処理]

まず、RNA二次構造予測処理の詳細について図7を参照して説明する。図7

は、本実施の形態における本システムのRNA二次構造予測処理の一例を示す処理概念図である。

【0079】

まず、文法データベース106bに既知のRNAの構造トポロジーを表す文法を集積する。そして、利用者が構造未知のRNA配列であってその二次構造を特定したいものを入力装置112を介してRNA配列解析装置100に入力すると（ステップSA-1）、構造予測部102aは、構文解析部102bの処理により、文法データベース106bから文法を取り出し（ステップSA-2）、RNA配列に対して各文法を適合して構文解析（パーズ）を行う（ステップSA-3）。ここで、利用者のRNA配列の入力は、RNA配列データベース106aから所望の配列を選択することにより入力してもよく、外部システム200の外部データベースから所望の配列を選択することにより入力してもよく、さらに、所望の配列を直接入力してもよい。

【0080】

そして、構造予測部102aは、適合度計算部102cの処理により、パーズが成功し導出された構文解析木について、例えば、ループや、塩基対とその他の二次構造要素のそれぞれの自由エネルギーの合計などを計算することにより求める平衡自由エネルギー（ ΔG ）などに基づいて適合度を求める。ここで、適合度計算方法は、例えば上述した文献1から3に示した方法のほか、従来のいずれの方法を用いてもよい。

【0081】

そして、構造予測部102aは、計算された適合度のうち予め定めた条件を満たす適合度である構文解析木を適合度が高い順にソートする（ステップSA-4）。

【0082】

そして、構造予測部102aは、入出力制御インターフェース部108を介して出力装置114にソートされた構文解析木とその適合度を出力することにより、利用者が入力した一配列に対して多文法で構文解析を行うことができるようになる。すなわち、各生成文法に対してそれぞれ構文解析し適合度計算を行い適合

度を得る。その結果、生成文法ごとに適合度が得られることになり、それらの適合度をソートすることによって生成文法に順位を付ける。これにより、生成文法に対する構造トポロジーにも順位が付けられることになるので、最終的にRNA配列が取り得る可能性の高い順に構造トポロジーを確認することができるようになる。これにて、RNA二次構造予測処理が終了する。

【0083】

[同一構造RNA配列抽出処理]

次に、同一構造RNA配列抽出処理の詳細について図8を参照して説明する。

図8は、本実施の形態における本システムの同一構造RNA配列抽出処理の一例を示す処理概念図である。

【0084】

まず、利用者は、特定の構造トポロジーに対応する文法を文法データベース106bから選択する。そして、構造予測部102aは、構文解析部102bの処理により、RNA配列データベース106aからRNA配列を取り出し（ステップSB-1）、各RNA配列に対して文法を適合して（ステップSB-2）、構文解析を行う（ステップSB-3）。

【0085】

そして、適合度計算部102cは、導出された構文解析木に対して適合度計算を行う。そして、構造予測部102aは、計算された適合度のうち予め定めた条件を満たす適合度である構文解析木を導出したRNA配列を、指定した文法が表す構造トポロジーの二次構造を持つRNA配列の候補として抽出する（ステップSB-4）。

【0086】

そして、構造予測部102aは、抽出されたRNA配列を当該文法が表す構造トポロジーの二次構造をもつ可能性のあるものとして、入出力制御インターフェース部108を介して出力装置114に出力する（ステップSB-5）。これにて、同一構造RNA配列抽出処理が終了する。

【0087】

[共通構造抽出処理]

次に、共通構造抽出処理の詳細について図9を参照して説明する。図9は、本実施の形態における本システムの共通構造抽出処理の一例を示す処理概念図である。

【0088】

まず、構造予測部102aは、RNA配列データベース106aから1つまたは2つ以上のRNA配列を取り出し（ステップSC-1およびステップSC-2）、構文解析部102bの処理により、各RNA配列に対して、文法データベース106bから取り出した（ステップSC-3）、1つまたは2つ以上の文法を適合する（ステップSC-4）。RNA配列解析装置100は、これらの取り出しや、パーズ処理について並列処理を行ってもよく、また、順次処理を行ってもよい。

【0089】

そして、適合度計算部102cは、導出された構文解析木に対して適合度計算を行い、共通構造マトリックス作成部102fの処理により、計算された適合度のうち予め定めた条件を満たす適合度である構文解析木を導出したRNA配列を抽出する（ステップSC-5）。

【0090】

そして、共通構造マトリックス作成部102fは、適合した文法が表す構造トポロジーとRNA配列とを二次元マトリックスで表示し、二次元マトリックスにおいて抽出されたRNA配列と構造トポロジーに対応する格子部分をマークすることにより、RNA配列間で共通に有する構造トポロジーを可視化する（ステップSC-6）。

【0091】

ここで、マークは、図9に示すように対象の格子部分に特定の色をつけてもよく、また、特定の記号（○など）や文字（「Y」など）を記載してもよい。これにより、例えば、縦方向にマークが連続した場合（図9に示す例においては、2番目の構造トポロジーの列）には、この構造トポロジーが各RNA配列に共通に持っている配列であることが可視的に把握することができるようになる。これにて、共通構造抽出処理が終了する。

【0092】

[構造類似度計算処理]

次に、構造類似度計算処理の詳細について図10を参照して説明する。図10は、本実施の形態における本システムの構造類似度計算処理の一例を示す処理概念図である。

【0093】

まず、利用者が、入力装置112により類似度を計算したい複数（図10の例では2個）のRNA配列をRNA配列解析装置100に入力する（ステップSE-1）。

【0094】

そして、類似度計算部102dは、文法データベース106bから1つまたは2つ以上の文法を取り出し（ステップSE-2）、構文解析部102bの処理により、入力したRNA配列について文法を適合して構文解析を行う（ステップSE-3）。また、適合度計算部102cは、導出された構文解析木に対して適合度計算を行う（ステップSE-4）。

【0095】

そして、類似度計算部102dは、文法を適合することにより導出された構文解析木と適合度（導出されなかった場合には、それを表現する特別な値を設定する）を各RNA配列ごとに対応付けてベクトル演算や内積の計算などをすることにより（ステップSE-5）、RNA配列間の類似度を計算する（ステップSE-6）。

【0096】

例えば、入力である*i*個のRNA配列をRNA₁, RNA₂, ..., RNA_iとし、文法データベース106bに格納されている*N*個の文法をG₁, G₂, ..., G_Nとし、また、RNA配列xと文法gに対するパーザが成功したときの適合度をr(x, g)とする。ここで、適合度は実数値とし、適合度が大きいほどその構造を取り易いことを示すものとする。

また、入力RNA_jに関する適合度のベクトルR_jにおいて、R_jのk番目の要素R_j[k]は、RNA_jとG_kに対するパーザが成功したときはr(RNA_j, G

k) とし、RNA_jとG_kに対するパーザが失敗したときは仮に「×」とする。

【0097】

このとき、類似度計算部102dによる類似度計算は、以下の手法により行われる。まず、2つのRNA配列の適合度のベクトルR₁とR₂を入力する。

【0098】

ついで、類似度計算部102dは、類似度ベクトルS₁、S₂とペナルティPを求める。ここで、「ペナルティP」は、R₁[k]とR₂[k]の片方だけが「パーザ失敗(×)」であるkの個数であり、「類似度ベクトルS₁、S₂」は、R₁[k]もR₂[k]も「パーザ失敗(×)」ではない箇所だけを抜き出したベクトルである。図12は、ペナルティPと類似度ベクトルS₁、S₂の概念を説明する図である。

【0099】

ついで、類似度計算部102dは、類似度ベクトルS₁、S₂間の距離Dを以下の方法により求める。まず、類似度ベクトルS₁、S₂の要素数(ベクトルの次元)をMとする。そして、類似度計算で一般的に用いられるユークリッド距離を用いて以下の数式により距離を計算する。

$$D = \sqrt{\sum \{(S_1[k] - S_2[k])^2\}}$$

($\sqrt{\sum}$ は平方根であり、 \sum は $k = 1 \sim M$ に関する総和である。)

【0100】

ここで、距離Dが大きい場合には類似度が低いことになり、また、ペナルティPが大きいと類似度が低いことになるので、ペナルティPと距離Dを用いて以下の数式により類似度Simを求める。

$$Sim = a^P / D$$

(aは定数($0 < a < 1$)である。)

【0101】

そして、Simを類似度として出力する。ここで、定数aを小さくすると、距離DよりもペナルティPが重視されることになる。これにて、構造類似度計算処理が終了する。

【0102】

【遺伝子予測処理】

次に、遺伝子予測処理の詳細について図11を参照して説明する。図11は、本実施の形態における本システムの遺伝子予測処理の一例を示す処理概念図である。

【0103】

まず、利用者が遺伝子部分が未知のDNA配列を入力装置112を介してRNA配列解析装置100に入力すると、RNA配列解析装置100は、遺伝子予測部102gの処理により、入力されたDNA配列に基づいて、当該DNA配列から転写されるRNA配列（以下、「予測RNA配列」という）を自動的に変換して作成する（ステップSF-1）。ここで、利用者のDNA配列の入力は、外部システム200の外部データベースやインハウスデータベースから所望のDNA配列を選択することにより入力してもよく、さらに、所望の配列を直接入力してもよい。

【0104】

ついで、構造予測部102aがこの予測RNA配列を構文解析部102bに入力すると（ステップSF-2）、構文解析部102bの処理により、文法データベース106bから1つまたは2つ以上の文法が取り出され（ステップSF-3）、各文法を予測RNA配列に適合する（ステップSF-4）。

【0105】

そして、適合度計算部102cは、構文解析部102bが導出した構文解析木について適合度計算を行い（ステップSF-5）、遺伝子予測部102gは、計算された適合度のうち予め定めた条件を満たす適合度である構文解析木を導出した予測RNA配列に対応するDNA配列部分を遺伝子の候補として予測する（ステップSF-6）。すなわち、DNA配列のうち、当該予測RNA配列の部分が遺伝子部分である可能性が高い領域として出力される。

【0106】

これにより、DNA配列のうち既知のトポロジーを有する可能性のある予測RNA配列に対応する部分について、遺伝子部分である可能性があることを予測することができるようになる。これにて、遺伝子予測処理が終了する。

【0107】

[実施例]

本発明の実施例について、以下に図13～図23を参照して説明する。

1 準備

本節では実施例の準備として、いくつかの具体的なRNA二次構造トポロジーを定義し、それらをモデリングする生成文法を規定する。本実施例では説明の便宜上、生成文法として文脈自由文法を用いるが、よりモデリング能力の高いRNA木文法（文献1）を用いた場合でも同様のことが説明できる。

【0108】

1. 1 二次構造トポロジー

図13に示される2つのRNA二次構造トポロジーを考える。

ステムループは、ステム（H（a））とヘアピンループ（L（a））から構成される。2並列ステムループは、並列に並んだ2つのステムループから構成される。それぞれのステム部分（H₁（b）、H₂（b））とヘアピンループ部分、（L₁（b）、L₂（b））の他にステムとステムをつなぐループ部分（I（b））がある。

【0109】

上記構造トポロジーについて、さらに具体的な特徴を考えることができる。例えばステムやループ部分のサイズの制約、ステムを構成する塩基対にミスマッチ（内部ループやバルジループ）を許すかどうか、あるいは特定の場所に特定の塩基配列を含むかどうか、などといったより詳細な特徴を持ったトポロジーを考えることが可能である。そこで、本実施例では次のような特徴を持ったRNA二次構造トポロジーT₁、T₂を扱う。

【0110】

トポロジーT₁

- 以下の特徴を持ったシステムループ構造（図13（a）参照）である。
- システム（H（a））を構成する塩基対はミスマッチを含まない。
- システム（H（a））サイズは1塩基対以上とする。
- ヘアピンループ（L（a））のサイズは1塩基以上とする。

【0111】

トポロジー T_2

- 以下の特徴を持った2並列システムループ構造（図13（b）参照）である。
- トポロジー T_1 を2つ並列に並べたもの。
- システム（ H_1 （b））とシステム（ H_2 （b））の間のループ（I（b））の長さは1塩基以上とする。

【0112】

1. 2 文脈自由文法による二次構造トポロジーのモデリング

以上のように定義された2つのトポロジー T_1 、 T_2 を文脈自由文法を用いてモデリングする。文脈自由文法は一般に以下の4項組によって規定される。

$$G = (N, \Sigma, P, S)$$

N は非終端記号の有限集合、 Σ は終端記号の有限集合、 P は生成規則の有限集合、 S は開始記号を表す。

【0113】

しかしながら、本実施例では常に $\Sigma = \{a, u, g, c\}$ 、開始記号は S 、さらに N は生成規則 P に出現する非終端記号のみからなるものとすることから、 P のみを指定することにより文脈自由文法 G を規定することができる。よって便宜上、本稿では文脈自由文法 G を規定する際には、生成規則の有限集合 P のみを指定することにする。

【0114】

（1）トポロジー T_1 は以下の生成規則からなる文脈自由文法 G_1 によってモデリングされる。

【数1】

$$\begin{aligned} S &\rightarrow xH\bar{x} \\ H &\rightarrow xH\bar{x} \mid L \\ L &\rightarrow xL \mid x \end{aligned}$$

ただし、 $x \in \Sigma$ とし、 \bar{x} は x と塩基対を形成する相手の塩基とする。

【0115】

すなわち、ワトソン-クリック塩基対のみを考える場合は、一番最初の生成規則は以下と同値である。

$$S \rightarrow aH u \mid uH a \mid gH c \mid cH g$$

非ワトソン-クリック塩基対を許す場合はさらに、 $S \rightarrow gH u$ などを追加してもよい。

【0116】

G_1 において、

【数2】

$$S \rightarrow xH\bar{x} \text{ と } H \rightarrow xH\bar{x}$$

によって塩基対（ステムを構成）が生成され、 $L \rightarrow xL$ と $L \rightarrow x$ によって塩基対を形成しない塩基（ループを構成）が生成されるとみなす。すると、 G_1 は RNA の二次構造を生成することができるうことになる。このように、任意の文脈自由文法 G に対して、 G が生成することができるすべての RNA 二次構造からなる集合 $SS(G)$ が規定されることになる。

【0117】

G_1 がトポロジー T_1 をモデリングする、とは以下が成り立つことを言う。“ G_1 はトポロジー T_1 の性質を満たすすべての RNA 二次構造を生成することができる”

、なおかつ、 G_1 で生成することができるすべてのRNA二次構造はトポロジー T_1 の性質を満たす。”

【0118】

これは、 G_1 による導出をみれば自明である。 G_1 による導出はすべて以下のようになる。

【数3】

ただし、 $n \geq 1$ 、 $l \geq 1$ とする。

$$\begin{aligned}
 S &\rightarrow x_1 H \overline{x_1} \\
 &\rightarrow x_1 x_2 H \overline{x_2} \overline{x_1} \rightarrow \dots \rightarrow x_1 x_2 \dots x_n H \overline{x_n} \dots \overline{x_2} \overline{x_1} \\
 &\rightarrow x_1 x_2 \dots x_n L \overline{x_n} \dots \overline{x_2} \overline{x_1} \\
 &\rightarrow x_1 x_2 \dots x_n y_1 L \overline{x_n} \dots \overline{x_2} \overline{x_1} \\
 &\rightarrow \dots \rightarrow x_1 x_2 \dots x_n y_1 \dots y_{l-1} L \overline{x_n} \dots \overline{x_2} \overline{x_1} \\
 &\rightarrow x_1 x_2 \dots x_n y_1 \dots y_{l-1} y_l \overline{x_n} \dots \overline{x_2} \overline{x_1}
 \end{aligned}$$

このとき、 $x_1 x_2 \dots x_n$ と、これに対応する $\overline{x_n} \dots \overline{x_2} \overline{x_1}$ の部分がシステムに、 $y_1 \dots y_{l-1} y_l$ の部分がヘアピンループに対応する。また、 $n \geq 1$ 、 $l \geq 1$ であるからシステムのサイズは1塩基対以上、ヘアピンループのサイズは1塩基以上である。

【0119】

よって、 G_1 は T_1 をモデリングすることがわかる。

【0120】

(2) トポロジー T_2 は以下の生成規則からなる文脈自由文法 G_2 によってモデリングされる。

【数4】

$$\begin{aligned}
 S &\rightarrow S_1 S_2 \\
 S_1 &\rightarrow x H \bar{x} \\
 S_2 &\rightarrow x H \bar{x} \\
 H &\rightarrow x H \bar{x} \mid L \\
 L &\rightarrow x L \mid x \\
 I &\rightarrow x L
 \end{aligned}$$

【0121】

以下の生成規則からなる文脈自由文法 G_0 は、文脈自由文法によって生成することができる可能なすべての RNA 二次構造を生成することができる万能な文脈自由文法である。

【数5】

$$S \rightarrow SS \mid x S \bar{x} \mid x S \mid S x \mid x \mid \lambda$$

ただし、 λ は空文字を表す。例えば、 G_1 によるいかなる導出も G_0 によってシミュレートできる。すなわち、以下のような導出を G_0 によって行なうことが可能である。

【数6】

$$\begin{aligned}
 S &\rightarrow x_1 S \bar{x_1} \\
 &\rightarrow x_1 x_2 S \bar{x_2} \bar{x_1} \rightarrow \dots \rightarrow x_1 x_2 \dots x_n S \bar{x_n} \dots \bar{x_2} \bar{x_1} \\
 &\rightarrow x_1 x_2 \dots x_n S \bar{x_n} \dots \bar{x_2} \bar{x_1} \\
 &\rightarrow x_1 x_2 \dots x_n y_1 S \bar{x_n} \dots \bar{x_2} \bar{x_1} \\
 &\rightarrow \dots \rightarrow x_1 x_2 \dots x_n y_1 \dots y_{l-1} S \bar{x_n} \dots \bar{x_2} \bar{x_1} \\
 &\rightarrow x_1 x_2 \dots x_n y_1 \dots y_{l-1} y_l \bar{x_n} \dots \bar{x_2} \bar{x_1}
 \end{aligned}$$

【0122】

上記の導出は非終端記号以外、すなわち生成される RNA 二次構造は、 G_1 によって生成されるものとまったく同じである。よって G_1 が生成可能なすべての二次構造を G_0 によって生成可能なことがわかる。すなわち、

$SS(G_0) \supseteq SS(G_1)$

である。

【0123】

このように、どんな文脈自由文法Gに対しても

$SS(G_0) \supseteq SS(G)$

が成り立つことが知られている。以降では、このような万能文法によって生成される二次構造全体を“すべての二次構造”と考える。

【0124】

1. 3 構文解析木と適合度

ある与えられたRNA配列がある与えられたRNA二次構造トポロジーの性質を満たす二次構造を形成できるかどうかという問題は、対象トポロジーをモデリングした文法によって対象配列が導出できるかという問題に対応する。これは生成文法の構文解析アルゴリズムによって解くことができる。

【0125】

構文解析アルゴリズムは、与えられた文法によって、与えられた配列が導出できるかどうかを判定し、導出可能な場合はその導出過程、すなわち構文解析木を出力する。二次構造トポロジーをモデリングした文法において、構文解析木は二次構造を表現しているので、構文解析アルゴリズムは、対象トポロジーに適合する具体的な二次構造を出力すると解釈してよいことになるからである。

【0126】

RNA配列 $s_1 = g g g g a a a a c c c c$ がトポロジー T_1, T_2 に適合する二次構造を形成できるかどうかについて考える。

配列 s_1 は G_1 によって以下のように導出できる。これにより配列 s_1 は T_1 に適合する二次構造をとりうることがわかる。

$S \rightarrow gHc \rightarrow ggHcc \rightarrow gggHccc \rightarrow ggggHcccc$
 $\rightarrow ggggLcccc \rightarrow ggggaLcccc \rightarrow ggggaalcccc$
 $\rightarrow ggggaaacccc$ (1)

【0127】

また、 s_1 は G_1 によって以下のようにも導出できる。

$S \rightarrow gHc \rightarrow ggHcc \rightarrow gggHccc$
 $\rightarrow gggLccc \rightarrow ggggLccc \rightarrow ggggaLccc$
 $\rightarrow ggggaalccc \rightarrow ggggaaaLccc$
 $\rightarrow ggggaaacccc$ (2)

【0128】

しかし、 s_1 は G_2 によって導出することはできない。これにより s_1 はトポロジー T_2 に適合する二次構造をとりえないことがわかる。

【0129】

s_1 を G_1 によって上記の2通りの方法で導出するとき、それぞれの導出に対応する構文解析木とそれが表現する二次構造を図14に示す。すなわち、(1)のように導出した場合は、図14の(1)に示される構文解析木と二次構造が生成され、(2)のように導出した場合は、図14の(2)に示される構文解析木と二次構造が生成される。

【0130】

この例のように複数の構文解析木が得られたときに、どの構文解析木、すなわち、どの二次構造を結果として出力するかを決定する必要が生じる。そのために、何らかの評価関数によって構文解析木（あるいは二次構造）にスコアを与え、構文解析木（あるいは二次構造）に順位を付ける必要が生じる。このようなスコアは文法によって異なる評価関数を用いても良いし、文法に依存しない絶対的な評価関数であってもよい。以降では、このスコアを適合度と呼ぶ。

【0131】

以下に、これまでに利用されている適合度の評価法の例を示すが、本発明によつて利用される適合度は以下のものに限定されない。

【0132】

(1) 塩基対数による適合度の評価

一般に、塩基対を形成する際の水素結合によってRNA分子はエネルギー的に安定なものになる。そこでこの評価法では、単純に塩基対の数が多い二次構造ほど優先させる。つまり、構文解析木の適合度として、対応する二次構造の塩基対数を用いる。この評価法において、上記の例の適合度を評価すると、図14の(1)に示される構文解析木は適合度3となり、(2)に示される構文解析木は適合度2となり、適合度の大きい(1)の構造が採用されることになる。

【0133】

本評価法に基づいた代表的な手法として、Nussinovの折り畳みアルゴリズム [Nussinov, R., Piecxenk, G., Geiggs, J. R., and Kleitman, D. J., "Algorithms for loop matchings," SIAM journal of Applied Mathematics, 35, 68-82, 1978] がある。

【0134】

(2) 平衡自由エネルギー(ΔG)による適合度の評価

RNA二次構造の物理化学的な安定度を計算するために、小さなモデルRNA分子の熱力学的な実験によって決定された平衡自由エネルギー(ΔG)パラメータを利用する方法がある。ある二次構造の(ΔG)は、それを構成する塩基対やループなどの二次構造要素に対する自由エネルギーの合計で近似される。この自由エネルギーパラメータでは、塩基対により構造が安定化し、ループにより構造が不安定化する。各二次構造要素の詳細なパラメータは [Turner, D. H., Sugimoto, N., Jaeger, J. A., Longfellow, C. E., Freier, S. M., and Kierzek, R., "Improved parameters for prediction of RNA structure," Cold Spring Harbor

Symposia Quantitative Biology, 52, 12
 3-133, 1987] に示されている。ここでは塩基対の自由エネルギーを図
 15に、ループの自由エネルギーを図16に示す。

【0135】

上記の自由エネルギーパラメータを用いて図14の構造(1)と(2)の(ΔG)を求めるとき、それぞれ以下のようになる。

$$\begin{aligned}\Delta G \text{ (構造 (1))} &= \Delta G (g c, g c) + \Delta G (g c, g c) \\ &\quad + \Delta G (g c, g c) \\ &\quad + (\Delta G) \text{ (サイズ3のヘアピンループ)} \\ &= (-2.9) + (-2.9) + (-2.9) \\ &\quad + 7.4 = -1.3\end{aligned}$$

$$\begin{aligned}\Delta G \text{ (構造 (2))} &= \Delta G (g c, g c) + \Delta G (g c, g c) \\ &\quad + \Delta G \text{ (サイズ5のヘアピンループ)} \\ &= (-2.9) + (-2.9) + 4.4 = -1.4\end{aligned}$$

【0136】

ここで注意すべきことは、塩基対の自由エネルギーの計算方法である。連続して積み重なった2組の塩基対に対してひとつのエネルギー値が与えられる。すなわち、構造(1)では5'側から数えて1番目のg c塩基対と2番目のg c塩基対に対して、 $\Delta G (g c, g c)$ が計算され、2番目のg c塩基対と3番目のg c塩基対に対して、 $\Delta G (g c, g c)$ が計算され、3番目のg c塩基対と4番目のg c塩基対に対して、 $\Delta G (g c, g c)$ が計算される。これに対し構造(2)では5'側から数えて1番目のg c塩基対と2番目のg c塩基対に対して、 $\Delta G (g c, g c)$ が計算され、2番目のg c塩基対と3番目のg c塩基対に対して、 $\Delta G (g c, g c)$ が計算される。

【0137】

構文解析木の適合度を $-\Delta G$ と定めると、(1)の適合度は1.3となり、(2)の適合度は1.4となり、結果として適合度の大きい(2)の構造が採用されることになる。

【0138】

ΔG に基づいた代表的なRNA二次構造予測システムとして、ZukerのModel（文献3）がある。

【0139】

(3) 導出確率による適合度の評価

確率文法とは個々の生成規則にその適用確率が付加された生成文法である。例えば G_1 の各生成規則に以下のような確率 p が付加されている確率文脈自由文法 G_1 を考える。

$$p(S \rightarrow aHu) = 0.2$$

$$p(S \rightarrow uHa) = 0.2$$

$$p(S \rightarrow gHc) = 0.3$$

$$p(S \rightarrow cHg) = 0.3$$

$$p(H \rightarrow aHu) = 0.2$$

$$p(H \rightarrow uHa) = 0.2$$

$$p(H \rightarrow gHc) = 0.3$$

$$p(H \rightarrow cHg) = 0.3$$

$$p(H \rightarrow L) = 0.1$$

$$p(L \rightarrow aL) = 0.2$$

$$p(L \rightarrow uL) = 0.2$$

$$p(L \rightarrow gL) = 0.15$$

$$p(L \rightarrow cL) = 0.15$$

$$p(L \rightarrow a) = 0.1$$

$$p(L \rightarrow u) = 0.1$$

$$p(L \rightarrow g) = 0.05$$

$$p(L \rightarrow c) = 0.05$$

【0140】

このとき G_1 による s_1 の導出確率は次のようにして計算される。すなわち、(

1) の導出確率は、

$$\begin{aligned}
 & p(S \rightarrow g H c) \times p(H \rightarrow g H c) \times p(H \rightarrow g H c) \times p(H \rightarrow g H c) \times \\
 & p(H \rightarrow L) \times p(L \rightarrow a L) \times p(L \rightarrow a L) \times p(L \rightarrow a) \\
 = & 0.3 \times 0.3 \times 0.3 \times 0.1 \times 0.2 \times 0.2 \times 0.1 \\
 = & 0.00000324
 \end{aligned}$$

と計算される。また、(2) の導出確率は、

$$\begin{aligned}
 & p(S \rightarrow g H c) \times p(H \rightarrow g H c) \times p(H \rightarrow g H c) \times p(H \rightarrow L) \times p(L \rightarrow g L) \\
 & \times p(L \rightarrow a L) \times p(L \rightarrow a L) \times p(L \rightarrow a L) \times p(L \rightarrow c) \\
 = & 0.3 \times 0.3 \times 0.3 \times 0.1 \times 0.15 \times 0.2 \times 0.2 \times 0.2 \\
 & \times 0.05 \\
 = & 0.000000162
 \end{aligned}$$

となる。

【0141】

そこで構文解析木の適合度として導出確率の自然対数をとると、(1) の適合度は $\ln 0.00000324 = -12.6$ 、(2) の適合度は $\ln 0.000000162 = -15.6$ となり、結果として適合度の大きい(1)の構造が採用されることになる。

【0142】

本評価法の根本である、各生成規則に付加されるべき確率パラメータは、最尤推定法と内側・外側アルゴリズム (inside-outside algorithm) などにより学習してもよいし、ヒューリスティクス (heuristics) などによって主観的に推定してもよい。例えば、文献 [Sakakibara "Stochastic Context-free Grammars for tRNA modeling," Nucleic Acids Research, 22, 5112-5120, 1994.] では複数の tRNA

A配列からtRNAの構造をモデリングする確率文脈自由文法を学習する手法について述べられている。

【0143】

以上いくつかの適合度評価法について説明したが、以降の説明では適合度として $-\Delta G$ を用いる。

【0144】

次にRNA配列 $s_2 = g c c c a u a g g c a a a g c c u a u g g g c$ がトポロジー T_1, T_2 に適合する二次構造を形成できるかどうかを考える。この場合も同様に s_2 が G_1, G_2 によって導出できるかどうかを調べればよい。結論からいうと、 s_2 は G_1, G_2 のいずれからも導出することができる。さらにどちらの文法でも複数の導出のしかたが存在する。それぞれの文法について $-\Delta G$ の適合度指標において最適な構文解析木とそれに対応する二次構造を図17に示す。

【0145】

それぞれの構造の ΔG を求めると、以下のようになる。

$$\begin{aligned}
 \Delta G \text{ (構造 (1))} &= \Delta G \text{ (g c, c g)} \times 2 + \Delta G \text{ (c g, c g)} \\
 &\quad \times 2 + \Delta G \text{ (c g, a u)} + \Delta G \text{ (a u, u a)} \\
 &\quad + \Delta G \text{ (u a, a u)} + \Delta G \text{ (a u, g c)} \\
 &\quad + \Delta G \text{ (g c, g c)} \\
 &\quad + \Delta G \text{ (サイズ3のヘアピンループ)} \\
 &= (-3.4) \times 2 + (-2.9) \times 2 + (-1.8) \\
 &\quad + (-0.9) + (-1.1) + (-1.7) \\
 &\quad + (-2.9) + 7.4 = -13.6
 \end{aligned}$$

$$\begin{aligned}
 \Delta G \text{ (構造 (2))} &= \Delta G \text{ (g c, c g)} \times 2 + \Delta G \text{ (c g, c g)} \times 2 \\
 &\quad + \Delta G \text{ (サイズ4のヘアピンループ)} \times 2 \\
 &= (-3.4) \times 2 + (-2.9) \times 2 + 5.9 \times 2 \\
 &= -6.7
 \end{aligned}$$

よってトポロジー T_1 に適合する RNA 二次構造において s_2 がとりうる最適な構造の適合度は 13.6 であることがわかる。また、トポロジー T_2 に適合する RNA 二次構造において s_2 がとりうる最適な構造の適合度は 6.7 であることがわかる。また、 s_2 を万能な文法 G_0 によって構文解析すると、最適構造として構造 (1) が見つかる。これにより、構造 (1) が“すべての二次構造”の中で最適な構造であることがわかる。このように万能文法によって構文解析を行なうことにより、すべての構造の中から最適な構造を見つけ出すことができる。

【0146】

本発明の基本となる“RNA配列を生成文法に適用して構文解析木を導出する構文解析手段と、上記構文解析手段にて導出された構文解析木に対して適合度の計算を行う適合度計算手段と、適合度最大の構文解析木に対応する二次構造を出力する最適二次構造出力手段”は、適合度計算を組み込んだ構文解析アルゴリズムにより実現されることになる。このような構文解析アルゴリズムを構造予測アルゴリズムと呼ぶ。 ΔG の適合度を指標にした RNA 木文法に基づく構造予測アルゴリズムは文献 1 に示されている。

【0147】

2. 発明実施例

本節では、前節までに定義した RNA 配列 s_1, s_2 、トポロジー T_1, T_2 およびそれらをモデリングする文脈自由文法 G_0, G_1, G_2 、さらに適合度として $-\Delta G$ を用いた場合の実施例について示す。

【0148】

はじめに、“RNA 二次構造の構造トポロジーと、当該構造トポロジーに適合する生成文法とを対応付けて格納する文法格納手段”においては、例えば (L e u-t RNA, G') や (16S r RNA, G'') などのようにある構造トポロジーに付けられた名称とその構造トポロジーをモデリングした文法が対応づけられて格納されている。本実施例では (ステムループ T_1, G_1) と (2 並列ステムループ T_2, G_2) を含むような文法 DB を仮定する。また、RNA 配列 s_1 と s_2 を含む RNA 配列 DB を仮定する。

【0149】

(1) 生成文法と適合度計算で構造候補を出力

あるRNA配列に対して、これがとりうる構造トポロジーを適合度が大きい順に知りたいとき、本発明によれば、以下の手順でこれを調べることができる。例として、入力配列を s_2 、検査対象トポロジー集合を T_1, T_2 とした場合について示す。

【0150】

手順1) RNA配列を配列DBから指定する。あるいは直接入力する。ここでは s_2 を指定する。

手順2) 検査対象とするトポロジーの集合（生成文法の集合）を文法DBから選択する。ここでは T_1 と T_2 (G_1 と G_2) を選択する。

手順3) 適合度の閾値を設定する。閾値は手順2で得られた各トポロジー（生成文法）に対してそれぞれ設定してもよいし、共通にひとつ設定してもよい。ここでは T_1 (G_1) に対して 10 を T_2 (G_2) に対して 5 を設定する。

手順4) 手順1で得られた配列を手順2で得られた各生成文法によってそれぞれ構文解析を行ない、適合度最大の構文解析木を求める。ここでは s_2 を G_1 によって構文解析し、最大の適合度 13.6 を持つ構文解析木を得る（図17(1) 参照）。

さらに s_2 を G_2 によって構文解析し、最大の適合度 6.7 を持つ構文解析木を得る（図17(2) 参照）。

手順5) 手順4で得られた構文解析木のうち手順3で得られた閾値以上の適合度をもつ構文解析木を適合度の大きい順にソートする。手順4で得られた G_1 に対する適合度 13.6 の構文解析木 1 は手順3で G_1 に対して設定された閾値 10 よりも大きいのでこれをソート対象とする。手順4で得られた G_2 に対する適合度 6.7 の構文解析木 2 は手順3で G_2 に対して設定された閾値 5 よりも大きいのでこれをソート対象とする。上でソート対象となった構文解析木を適合度の大きい順にソートすることによって、構文解析木 1、構文解析木 2 の順に順序づけされる。

手順6) 手順5でソートされた構文解析木の順に、対応するトポロジー名、適合度、構文解析木（二次構造）などを出力する。構文解析木 1 に対応して、ステ

ループ T_1 、適合度 13.6、図 17 (1) に示された二次構造を出力する。
 構文解析木 2 に対して、2 並列システムループ T_2 、適合度 6.7、図 17 (2)
) に示された二次構造を出力する。

以上の結果から、選択されたトポロジー集合のなかで s_2 が適合する構造候補
 が図 18 のように出力される。

【0151】

従来の二次構造予測プログラムでは、与えられた配列がとりうる構造のなかで
 最適あるいは準最適な二次構造を順に出力するだけで、出力された構造がどのよ
 うなトポロジーであるかはユーザが判断しなければならなかった。本発明によれば、
 構造とトポロジーとを対応付けて出力することができるので、予測結果の確
 認にかかる労力が大きく軽減されることが期待される。

【0152】

また、本発明の実施について厳密に上記の手順と同じである必要はない。例え
 ば、手順 1 と 2 は順序をいれかえてもよいし、手順 5 の閾値により構文解析木を
 取捨選択する部分は手順 4 の構文解析部分に含めてしまってもよい。

【0153】

(2) 同じ構造を持つ配列の候補を出力

ある構造トポロジーに対して、これに適合する二次構造をとりうる RNA 配列
 を探したいとき、本発明によれば、以下の手順でこれを調べることができる。例
 として、入力構造トポロジーを T_2 、検査対象配列集合を s_1 、 s_2 とした場合に
 ついて示す。

手順 1) トポロジー (生成文法) を文法 DB から選択する。ここでは T_2 (G_2)
 を選択する。

手順 2) 適合度の閾値を設定する。ここでは 5 を選択する。

手順 3) 検査対象とする RNA 配列集合を配列 DB から選択する。あるいは直
 接入力する。ここでは s_1 、 s_2 選択する。

手順 4) 手順 3 で得られた各配列を、手順 1 で得られた生成文法によってそれ
 ぞれ構文解析を行ない、適合度最大の構文解析木をそれぞれ求める。ここでは s_1
 を G_2 によって構文解析し、導出不能であることを得る。さらに s_2 を G_2 によっ

て構文解析し、最大の適合度 6. 7 を持つ構文解析木を得る。（図 17 (2) 参照）

手順 5) 手順 4 で得られた構文解析木のうち手順 2 で得られた閾値以上の適合度をもつ構文解析木に対応する配列を出力する。手順 4 で得られた s_2 の G_2 に対する適合度 6. 7 の構文解析木は手順 2 で設定された閾値 5 よりも大きいので s_2 を出力する。以上の結果から、選択されたトポロジーをとりうる配列の候補が図 19 のように出力される。

【0154】

本発明の実施について厳密に上記の手順と同じである必要はない。例えば、手順 1 と 2 と 3 は任意の順にいれかえてもよいし、手順 5 は手順 4 の構文解析部分に含めてしまってもよい。

【0155】

(3) 共通構造抽出

ある RNA 配列の集合に対して、これらの配列が共通してとりうる構造トポロジーを調べたいとき、本発明によれば、以下の手順でこれを調べることができる。例として、入力配列集合を s_1 、 s_2 とし、検査対象構造トポロジーの集合を T_1 、 T_2 とした場合について示す。

【0156】

手順 1) RNA 配列の集合を配列 DB から指定する。あるいは直接入力する。ここでは s_1 と s_2 を指定する。

手順 2) 検査対象とするトポロジーの集合（生成文法の集合）を文法 DB から選択する。ここでは T_1 (G_1) と T_2 (G_2) を選択する。

手順 3) 適合度の閾値を設定する。閾値は手順 2 で得られた各トポロジー（生成文法）に対してそれぞれ設定してもよいし、共通にひとつ設定してもよい。ここでは共通に 0 を設定する。

手順 4) 手順 1 で得られた各配列を、手順 2 で得られた各生成文法によってそれぞれ構文解析を行ない、適合度最大の構文解析木を求める。

s_1 を G_1 によって構文解析し、最大の適合度 1. 4 を持つ構文解析木を得る（図 14 (2) 参照）。

s_1 を G_2 によって構文解析し、導出不能であることを得る。

s_2 を G_1 によって構文解析し、最大の適合度13.6を持つ構文解析木を得る。
(図17(1)参照)

s_2 を G_2 によって構文解析し、最大の適合度6.7を持つ構文解析木を得る。

(図17(2)参照)

手順5) 手順4で得られた構文解析木のうち閾値以上の適合度を持つ構文解析木を抽出する。手順4で得られたすべての構文解析木は手順3で得られた閾値0よりも大きい適合度を持つので手順4で得られたすべての構文解析木を抽出する。

手順6) 手順1で得られた配列集合を行に、手順2で得られたトポロジー集合を列に、手順5で得られた構文解析木の適合度を要素を持つマトリックスを作成する。図20に示すマトリックスを得る。

以上の結果得られたマトリックスを出力すれば、対象配列集合が共通してとりうる構造トポロジーを容易に確認することが可能になる。あるいは、以下の追加手順を実行すれば、共通構造の候補を順位づけて出力することができる。

手順7) 手順6で得られたマトリックスの各列、すなわちトポロジー、に対してスコアを計算する。例えば、有効な行要素の数を各列ごとに計算しスコアとすると、 T_1 のスコアは2、 T_2 のスコアは1となる。例えば、各行の適合度の総和を各列ごとに計算しスコアとすると、 T_1 のスコアは15.0、 T_2 のスコアは6.7となる。

手順8) 手順7で得られたスコアの高い順にトポロジーをソートし、出力する。上記のいずれのスコアを用いても T_1 、 T_2 の順に出力される。

【0157】

また、本発明の実施について厳密に上記の手順と同じである必要はない。例えば、手順1と2は順序をいれかえてもよいし、手順5を手順4の構文解析部分に含めてしまってもよい。

【0158】

(4) ジーン・ファインダ

RNA遺伝子部分に対応する配列は、非常に安定な構造をとりやすいので、適

合度が高くなる。そこで本発明では、万能文法を用いて構文解析を行ない、適合度の高い配列を配列DBから選び出して遺伝子候補として出力する。例として、配列集合を s_1, s_2 とした場合について示す。

【0159】

手順1) 検査対象とするRNA配列の集合を配列DBから指定する。あるいは直接入力する。ここでは s_1 と s_2 を指定する。

手順2) 適合度の閾値を設定する。ここでは10を設定する。

手順3) 手順1で得られた各配列を万能文法 G_0 によってそれぞれ構文解析を行ない、適合度最大の構文解析木を求める。

s_1 を G_0 によって構文解析し、最大の適合度1.4を持つ構文解析木を得る。

s_2 を G_0 によって構文解析し、最大の適合度13.6を持つ構文解析木を得る

手順4) 手順3で得られた構文解析木のうち閾値以上の適合度をもつ構文解析木に対応する配列を遺伝子候補として出力する。手順3で得られた s_1 の構文解析木は閾値10に満たないので s_1 は出力しない。手順3で得られた s_2 の構文解析木は閾値10よりも大きいので s_2 を遺伝子候補として出力する。

本発明の実施について厳密に上記の手順と同じである必要はない。例えば、手順1と2は順序をいれかえてもよいし、手順4は手順3の構文解析部分に含めてしまってもよい。

【0160】

(5) RNA配列から同じ構造を持つRNA配列を出力

あるRNA配列集合に対して、これらと同じトポロジーをとりうるRNA配列を調べたいとき、(3)の発明と(2)の発明とを組み合わせた本発明によれば、以下の手順でこれを調べることができる。例として、入力配列を $s = g\text{ccc}$ 、以下手順でこれを調べることができる。例として、入力配列を $s = g\text{ccc}$ 、以下手順でこれを調べることができる。例として、入力配列を $s = g\text{ccc}$ 、検査対象トポロジー集合を T_1, T_2 、検査対象配列集合を s_1, s_2 とした場合について示す。

【0161】

手順1) RNA配列集合を入力する。ここでは s のみからなる配列集合を入力する。

手順2) 検査対象とするRNA配列の集合を配列DBから指定する。ここでは s_1 と s_2 を指定する。

手順3) 検査対象とするトポロジーの集合(生成文法の集合)を文法DBから選択する。ここでは $T_1 (G_1)$ と $T_2 (G_2)$ を選択する。

手順4) 適合度の閾値を設定する。閾値は手順3で得られた各トポロジー(生成文法)に対してそれぞれ設定してもよいし、共通にひとつ設定してもよい。ここでは共通に5を設定する。

手順5) 手順1で得られた各RNA配列を、手順2で得られた各生成文法によってそれぞれ構文解析を行ない、適合度最大の構文解析木をそれぞれ求める。ここでは s を G_1 によって構文解析し、最大の適合度3.1を持つ構文解析木を得る。図21(1)にこの構文解析木が表現する二次構造を示す。さらに s を G_2 によって構文解析し、最大の適合度5.1を持つ構文解析木を得る。図21(2)にこの構文解析木が表現する二次構造を示す。

手順6) 手順5で得られた構文解析木のうち、手順4で得られた閾値以上の適合度をもつ構文解析木に対応する構文解析木を抽出する。手順5で得られた構文解析木のうち、 G_2 で構文解析することによって得られた適合度5.1の構文解析木が手順4で得られた閾値5よりも大きいのでこれを抽出する。

手順7) 手順1で得られた配列集合を行に、手順3で得られたトポロジー集合を列に、手順6で得られた構文解析木の適合度を要素を持つマトリックスを作成する。図22に示すマトリックスを得る。

手順8) 手順6で得られたマトリックスの各列、すなわちトポロジー、に対してスコアを計算し、スコアの順にトポロジーをソートする。ここでは行の総和を各列ごとに計算しスコアとするが、1行しかないので結果として、 T_1 のスコアが未定義、 T_2 のスコアが5.1になる。スコアを持つものだけでソートすると、 T_2 のみが得られる。

手順9) 手順2で得られた各配列を、手順8で得られたトポロジーの順にそれぞれ対応する文法で構文解析を行ない、適合度最大の構文解析木をそれぞれ求める。ここでは s_1 を G_2 によって構文解析し、導出不能であることを得る。

さらに s_2 を G_2 によって構文解析し、最大の適合度6.7を持つ構文解析木を得る。

得る。（図17（2）参照）

手順10） 手順9で得られた構文解析木のうち手順4で得られた閾値以上の適合度をもつ構文解析木に対応する配列を出力する。このとき、あわせてトポロジーとそのトポロジーに対する手順8で得られたスコアを出力する。手順9で得られた s_2 の G_2 に対する構文解析木の適合度 6.7 は手順4で得られた閾値 5 よりも大きいので s_2 を出力する。あわせて、 T_2 とそのスコア 5.1 を出力する。

以上の結果から図23に示すような出力が得られる。

この結果、 s_2 がトポロジー T_2 において、 s と共通な構造をとりうることがわかるようになる。

【0162】

本発明の実施について厳密に上記の手順と同じである必要はない。例えば、手順1と2と3は任意の順に入れ換えてよいし、手順6は手順5の構文解析部分に含めてしまってもよいし、手順10の閾値により構文解析木を取捨選択する部分は手順9の構文解析部分に含めてしまってもよい。

【0163】

【他の実施の形態】

さて、これまで本発明の実施の形態について説明したが、本発明は、上述した実施の形態以外にも、上記特許請求の範囲に記載した技術的思想の範囲内において種々の異なる実施の形態にて実施されてよいものである。

【0164】

例えば、RNA配列解析装置100がスタンドアローンの形態でRNA配列解析方法を行う場合を一例に説明したが、RNA配列解析装置100とは別筐体で構成されるクライアント端末からの要求に応じてRNA配列解析方法を行い、その処理結果を当該クライアント端末に返却するように構成してもよい。

【0165】

また、構造予測部102aは、適合度計算部102cによる適合度計算を行いながら構文解析部102bにより構文解析木を導出してもよい。すなわち、構文解析木を導出する構文解析部102bと、導出された構文解析木に対して適合度の計算を行う適合度計算部102cをひとつのアルゴリズムにて実現してもよい。

。このように構成することにより、RNA配列と木文法に対して可能な構文解析木は無数（配列長に対して指数のオーダー）に存在するため、構文解析木を導出してから適合度計算を行いソートすると指数オーダーの計算時間と記憶容量が必要となるという問題点を解決することができる。

【0166】

また、実施の形態において説明した各処理のうち、自動的に行なわれるものとして説明した処理の全部または一部を手動的に行なうこともでき、あるいは、手動的に行なわれるものとして説明した処理の全部または一部を公知の方法で自動的に行なうこともできる。

特に、構造予測部102aは複数のタスクとして実現してもよく、それぞれのタスクで並列処理を行うように実現してもよい。

この他、上記文書中や図面中で示した処理手順、制御手順、具体的な名称、各種の登録データや検索条件等のパラメータを含む情報、画面例、データベース構成については、特記する場合を除いて任意に変更することができる。

【0167】

また、RNA配列解析装置100に関して、図示の各構成要素は機能概念的なものであり、必ずしも物理的に図示の如く構成されていることを要しない。

例えば、RNA配列解析装置100の各サーバが備える処理機能、特に制御部にて行なわれる各処理機能については、その全部または任意の一部を、CPU（Central Processing Unit）および当該CPUにて解釈実行されるプログラムにて実現することができ、あるいは、ワイヤードロジックによるハードウェアとして実現することも可能である。なお、プログラムは、後述する記録媒体に記録されており、必要に応じてRNA配列解析装置100に機械的に読み取られる。

【0168】

記憶部106に格納される各種のデータベース等（RNA配列データベース106a～共通構造マトリックス106c）は、RAM、ROM等のメモリ装置、ハードディスク等の固定ディスク装置、フレキシブルディスク、光ディスク等のストレージ手段であり、各種処理やウェブサイト提供に用いる各種のプログラム

やテーブルやファイルやデータベースやウェブページ用ファイル等を格納する。

【0169】

また、RNA配列解析装置100は、既知のパーソナルコンピュータ、ワークステーション等の情報処理端末等の情報処理装置にプリンタやモニタやイメージスキャナ等の周辺装置を接続し、該情報処理装置に本発明の方法を実現させるソフトウェア（プログラム、データ等を含む）を実装することにより実現してもよい。

【0170】

さらに、RNA配列解析装置100の分散・統合の具体的形態は図示のものに限られず、その全部または一部を、各種の負荷等に応じた任意の単位で、機能的または物理的に分散・統合して構成することができる。例えば、各データベースを独立したデータベース装置として独立に構成してもよく、また、処理の一部をCGI (Common Gateway Interface) を用いて実現してもよい。

【0171】

また、本発明にかかるプログラムを、コンピュータ読み取り可能な記録媒体に格納することもできる。ここで、この「記録媒体」とは、フレキシブルディスク、光磁気ディスク、ROM、EPROM、EEPROM、CD-ROM、MO、DVD等の任意の「可搬用の物理媒体」や、各種コンピュータシステムに内蔵されるROM、RAM、HD等の任意の「固定用の物理媒体」、あるいは、LAN、WAN、インターネットに代表されるネットワークを介してプログラムを送信する場合の通信回線や搬送波のように、短期にプログラムを保持する「通信媒体」を含むものとする。

【0172】

また、「プログラム」とは、任意の言語や記述方法にて記述されたデータ処理方法であり、ソースコードやバイナリコード等の形式を問わない。なお、「プログラム」は必ずしも单一的に構成されるものに限られず、複数のモジュールやライブラリとして分散構成されるものや、OS (Operating System) に代表される別個のプログラムと協働してその機能を達成するものをも含む

。なお、実施の形態に示した各装置において記録媒体を読み取るための具体的な構成、読み取り手順、あるいは、読み取り後のインストール手順等については、周知の構成や手順を用いることができる。

【0173】

また、ネットワーク300は、RNA配列解析装置100と外部システム200とを相互に接続する機能を有し、例えば、インターネットや、インターネットや、LAN(有線／無線の双方を含む)や、VANや、パソコン通信網や、公衆電話網(アナログ／デジタルの双方を含む)や、専用回線網(アナログ／デジタルの双方を含む)や、CATV網や、IMT2000方式、GSM方式またはPDC／PDC-P方式等の携带回線交換網／携帯パケット交換網や、無線呼出網や、Bluetooth等の局所無線網や、PHS網や、CS、BSまたはISDB等の衛星通信網等のうちいずれかを含んでもよい。すなわち、本システムは、有線・無線を問わず任意のネットワークを介して、各種データを送受信することができる。

【0174】

【発明の効果】

以上詳細に説明したように、本発明によれば、RNA二次構造の構造トポジーと、当該構造トポジーに適合する生成文法とを対応付けて格納し、RNA配列を生成文法に適用して構文解析木を導出し、導出された構文解析木に対して適合度を計算し、計算された適合度のうち予め定めた条件を満たす適合度である構文解析木を適合度が高い順にソートし、ソートされた構文解析木をRNA配列の二次構造の候補として出力するので、一配列に対して多文法で構文解析を行うことができるようになる。すなわち、各生成文法に対してそれぞれ構文解析し適合度計算を行い適合度を得る。その結果、生成文法ごとに適合度が得られることになり、それらの適合度をソートすることによって生成文法に順位を付ける。これにより、生成文法に対する構造トポジーにも順位が付けられることになるので、最終的にRNA配列が取り得る可能性の高い順に構造トポジーを確認することができるRNA配列解析装置、RNA配列解析方法、プログラム、および、記録媒体を提供することができる。

【0175】

また、本発明によれば、RNA二次構造の構造トポロジーと、当該構造トポロジーに適合する生成文法とを対応付けて格納し、RNA配列を生成文法に適用して構文解析木を導出し、導出された構文解析木に対して適合度の計算を行い、計算された適合度のうち予め定めた条件を満たす適合度である構文解析木を導出したRNA配列を上記構造トポロジーの二次構造を持つRNA配列の候補として出力するので、多配列に対して一文法で構文解析を行うことができるようになる。すなわち、与えられた特定の構造トポロジーに対し、対応する生成文法を取得し、これを用いてRNA配列データベースに格納されているすべてまたは一部のRNA配列をそれぞれ構文解析し、ある閾値以下の適合度で構文解析に成功したRNA配列群を結果として出力する。これにより、与えられた特定の二次構造を取り得るようなRNA配列を検索することができるRNA配列解析装置、RNA配列解析方法、プログラム、および、記録媒体を提供することができる。

【0176】

また、本発明によれば、RNA二次構造の構造トポロジーと、当該構造トポロジーに適合する生成文法とを対応付けて格納し、RNA配列を生成文法に適用して構文解析木を導出し、導出された構文解析木に対して適合度の計算を行い、計算された適合度のうち予め定めた条件を満たす適合度である構文解析木を導出したRNA配列を抽出し、構造トポロジーとRNA配列とを二次元マトリックスで表示し、二次元マトリックスにおいて抽出されたRNA配列と構造トポロジーに対応する格子部分をマークすることにより、RNA配列間で共通に有する構造トポロジーを可視化するので、RNA配列間の共通構造を容易に発見することができるRNA配列解析装置、RNA配列解析方法、プログラム、および、記録媒体を提供することができる。

【0177】

また、本発明によれば、RNA二次構造の構造トポロジーと、当該構造トポロジーに適合する生成文法とを対応付けて格納し、利用者が入力したDNA配列から転写されるRNA配列を作成し、作成されたRNA配列に対して生成文法を適用して構文解析木を導出し、導出された構文解析木に対して適合度の計算を行い

、計算された適合度のうち予め定めた条件を満たす適合度である構文解析木を導出したRNA配列に対応するDNA配列部分を遺伝子の候補として予測するので、DNA配列のうち既知のトポロジーを有する可能性のあるRNA配列に対応する部分について、遺伝子部分である可能性があることを予測することができるRNA配列解析装置、RNA配列解析方法、プログラム、および、記録媒体を提供することができる。

【0178】

また、本発明によれば、RNA二次構造の構造トポロジーと、当該構造トポロジーに適合する生成文法とを対応付けて格納し、RNA配列を生成文法に適用して構文解析木を導出し、導出された構文解析木に対して適合度の計算を行い、計算された適合度に基づいてRNA配列間の類似度を計算するので、RNA構造の類似度を容易に求めることができるRNA配列解析装置、RNA配列解析方法、プログラム、および、記録媒体を提供することができる。

【0179】

さらに、本発明によれば、RNA二次構造の構造トポロジーと、当該構造トポロジーに適合する生成文法とを対応付けて格納し、RNA配列を生成文法に適用して構文解析木を導出し、導出された構文解析木に対して適合度の計算を行い、計算された適合度のうち予め定めた条件を満たす適合度である構文解析木を導出したRNA配列を抽出し、構造トポロジーとRNA配列とを二次元マトリックスで表示し、二次元マトリックスにおいて抽出されたRNA配列と構造トポロジーに対応する格子部分に適合度を表示する適合度マトリックスを作成し、適合度マトリックスについて、適合度により構造トポロジーをソートし、他のRNA配列について当該ソートされた構造トポロジーの順番に対応する生成文法により構文解析を行い適合度が最大となる構文解析木を求め、予め定めた条件を満たす適合度を持つ構文解析木に対応する他のRNA配列を抽出するので、共通の構造を持つRNA配列を容易に発見することができるRNA配列解析装置、RNA配列解析方法、プログラム、および、記録媒体を提供することができる。

【図面の簡単な説明】

【図1】

RNAの構造トポロジーの一例を説明する図である。

【図2】

従来技術によるRNAの二次構造がシステムループをとる場合の構文解析木の一例を示す図である。

【図3】

従来技術によるRNA二次構造の構造トポロジーについて、文法が固定されるとそれに対応した構造トポロジーが規定されることを示す概念図である。

【図4】

従来技術である木文法パーザを用いて、既知の文法からRNA配列の構文解析木を導出する場合の一例を示す図である。

【図5】

本発明が適用される本システムの構成の一例を示すブロック図である。

【図6】

文法データベース106bに格納される情報の一例を示す図である。

【図7】

本実施の形態における本システムのRNA二次構造予測処理の一例を示す処理概念図である。

【図8】

本実施の形態における本システムの同一構造RNA配列抽出処理の一例を示す処理概念図である。

【図9】

本実施の形態における本システムの共通構造抽出処理の一例を示す処理概念図である。

【図10】

本実施の形態における本システムの構造類似度計算処理の一例を示す処理概念図である。

【図11】

本実施の形態における本システムの遺伝子予測処理の一例を示す処理概念図である。

【図12】

ペナルティ P と類似度ベクトル s_1 、 s_2 の概念を説明する図である。

【図13】

RNA二次構造トポロジーの例を示す図である。

【図14】

s_1 の構文解析木と二次構造を示す図である。

【図15】

塩基対の自由エネルギーを示す図である。

【図16】

ループの自由エネルギーを示す図である。

【図17】

それぞれの文法について $-\Delta G$ の適合度指標において最適な構文解析木とそれに対応する二次構造を示す図である。

【図18】

選択されたトポロジー集合のなかで s_2 が適合する構造候補を示す図である。

【図19】

選択されたトポロジーをとりうる配列の候補を示す図である。

【図20】

構文解析木の適合度を要素に持つマトリックスを示す図である。

【図21】

s の最適な二次構造を示す図である。

【図22】

構文解析木の適合度を要素に持つマトリックスを示す図である。

【図23】

出力結果の一例を示す図である。

【符号の説明】

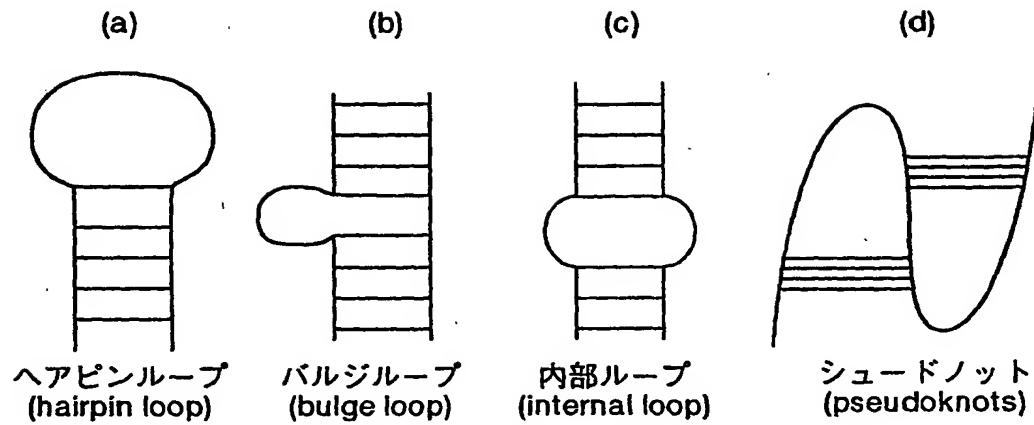
100 RNA配列解析装置

102 制御部

102a 構造予測部

102b 構文解析部
102c 適合度計算部
102d 類似度計算部
102f 共通構造マトリックス作成部
102g 遺伝子予測部
104 通信制御インターフェース部
106 記憶部
106a RNA配列データベース
106b 文法データベース
106c 共通構造マトリックス
108 入出力制御インターフェース部
112 入力装置
114 出力装置
200 外部システム
300 ネットワーク

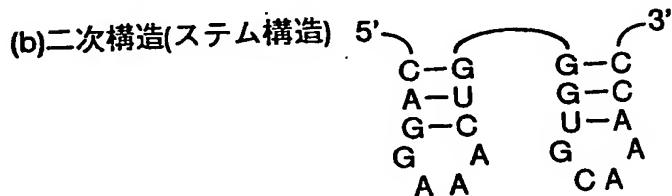
【書類名】 図面
【図1】



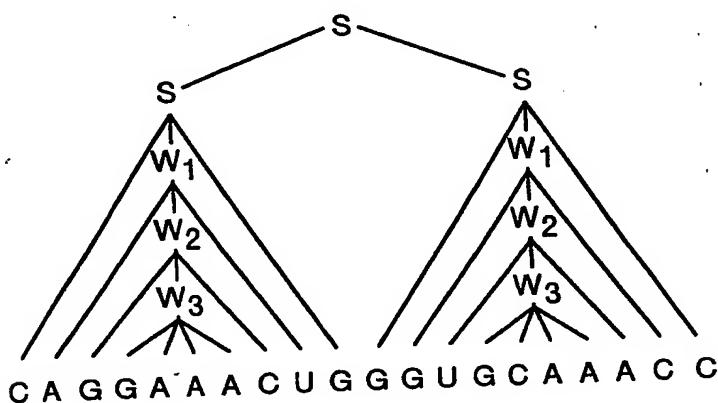
【図2】

(a)RNA配列

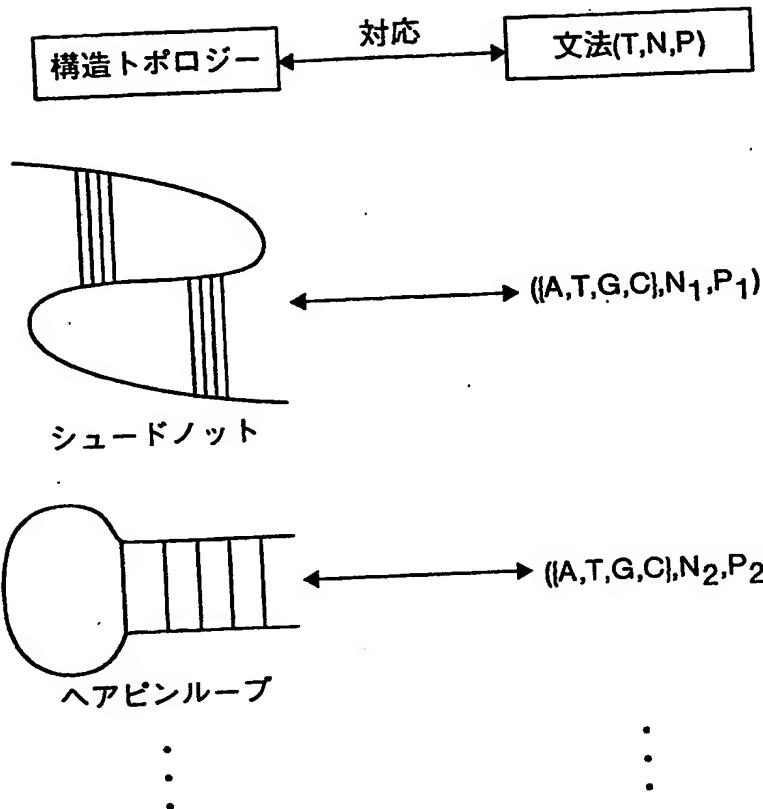
C A G G A A A C U G G G U G C A A A C C



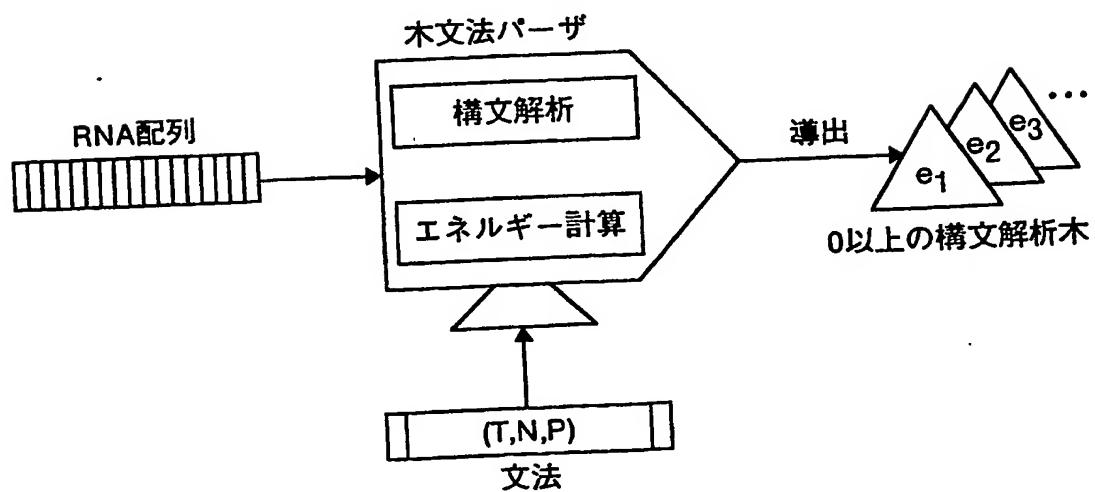
(c)構文解析木



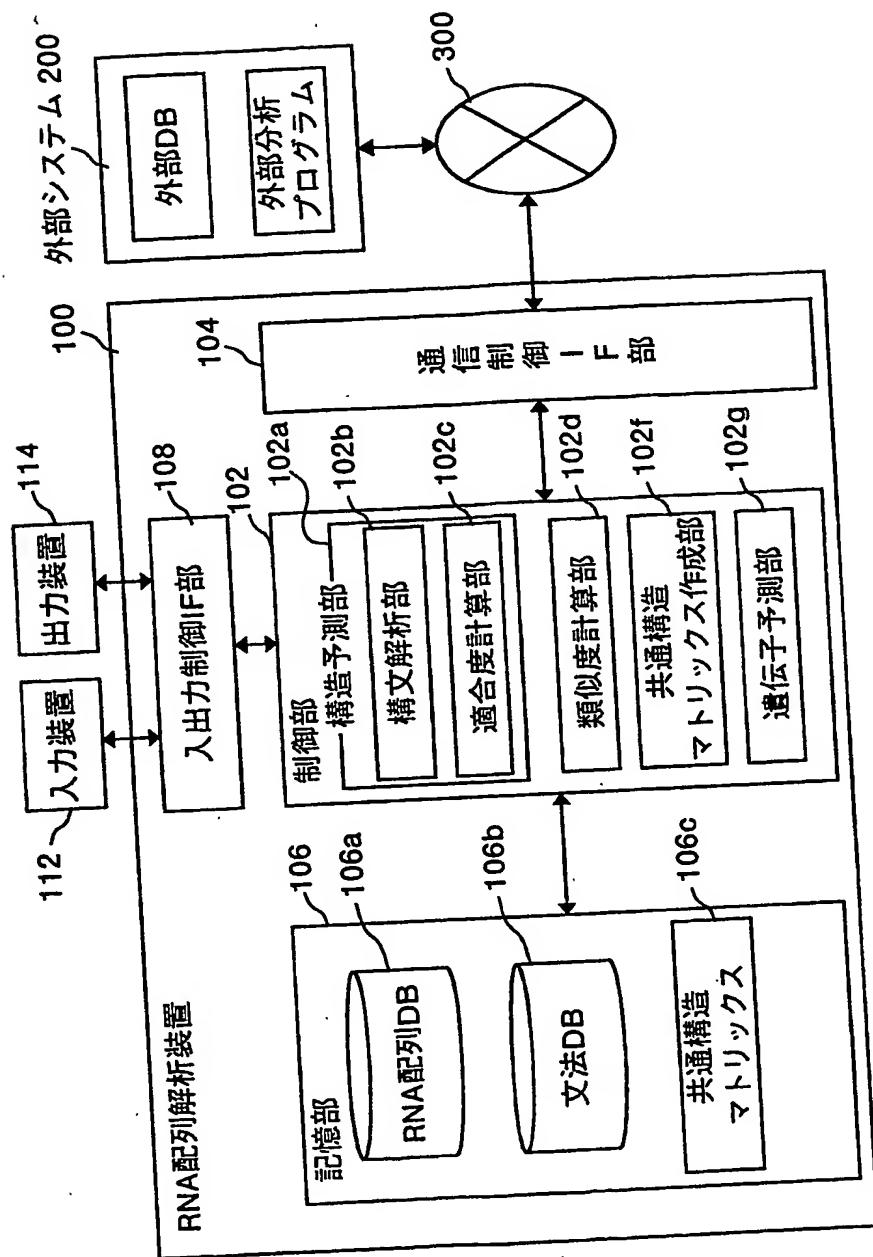
【図3】



【図4】



【図5】

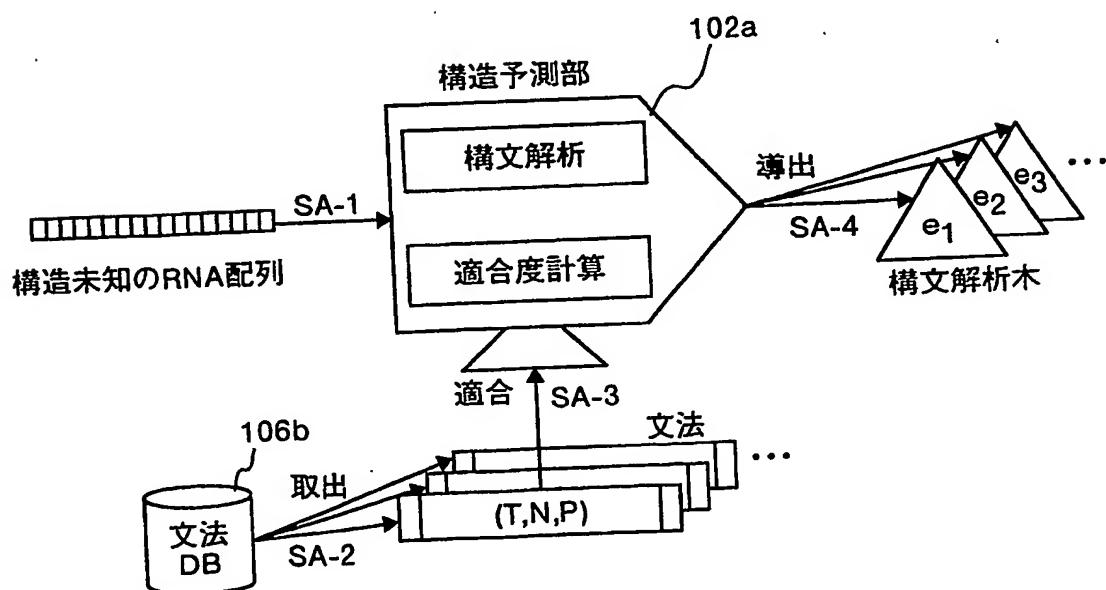


【図6】

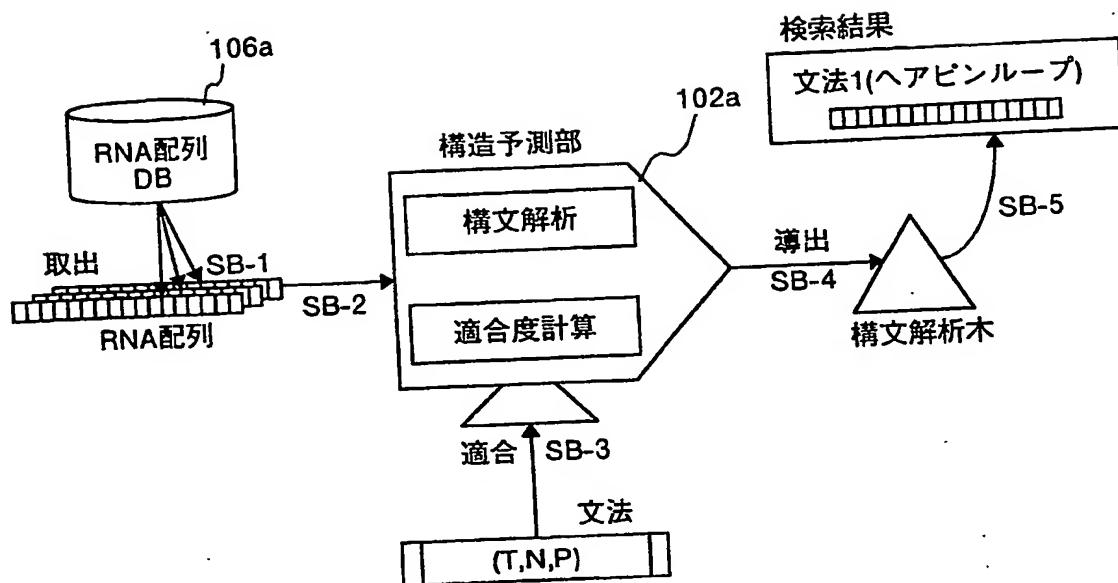
文法データベース 106b

構造トポロジー	文法		
	終端記号T	非終端記号N	生成規則P
シュードノット	{A,T,G,C}	N_1	P_1
ヘアピンループ		N_2	P_2
内側ループ		N_3	P_3
多枝ループ		N_4	P_4
バルジループ		N_5	P_5
⋮		⋮	⋮

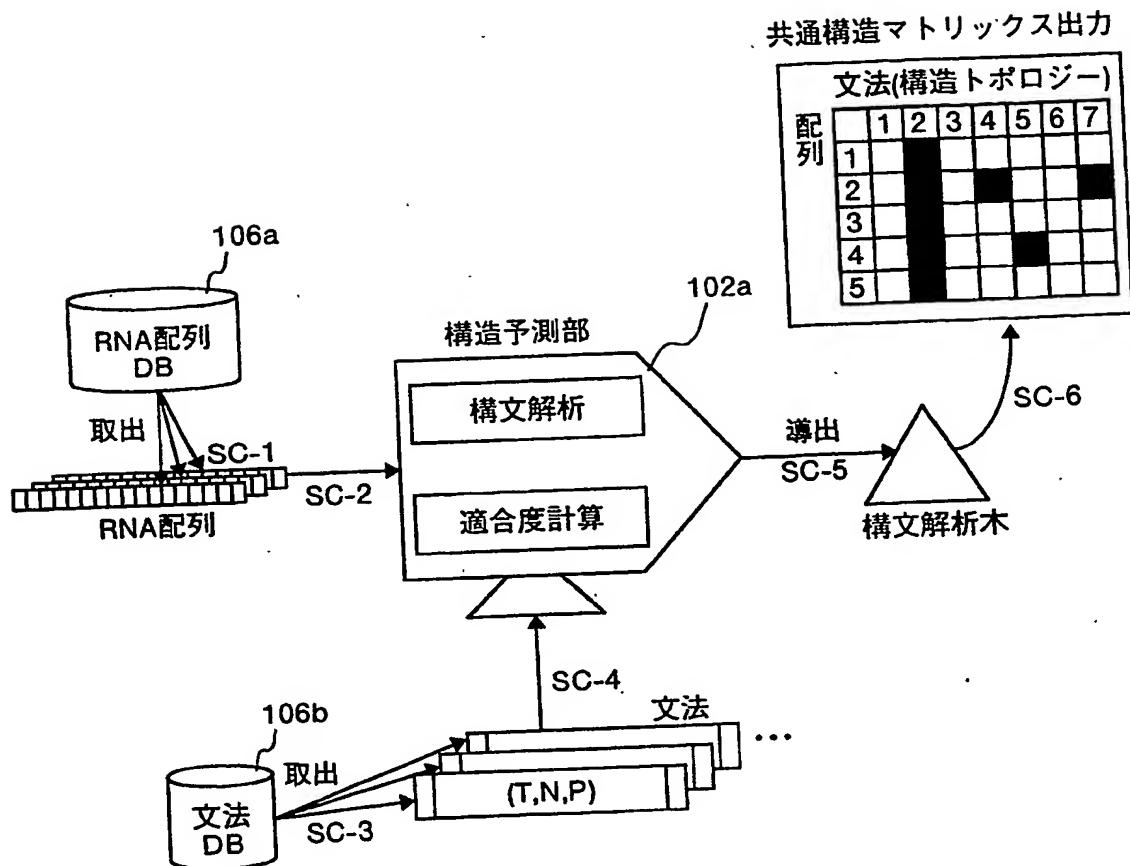
【図7】



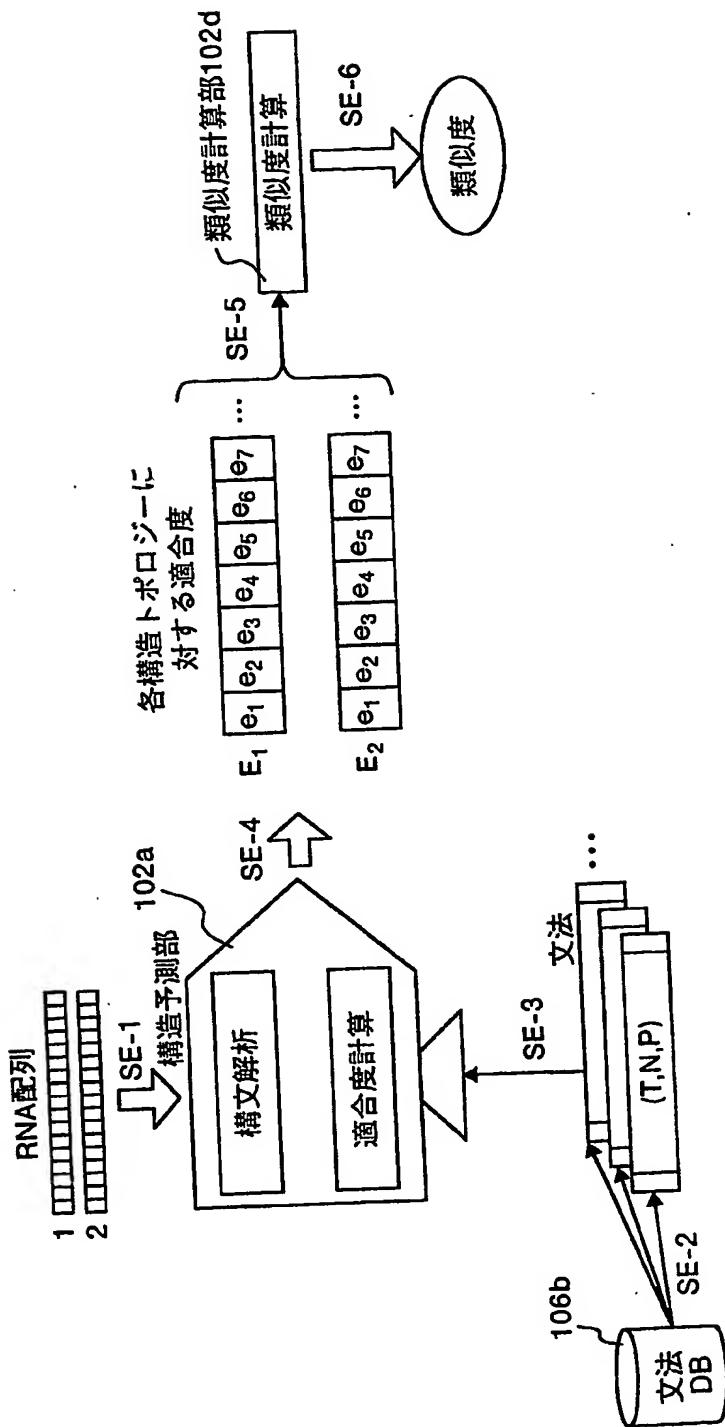
【図8】



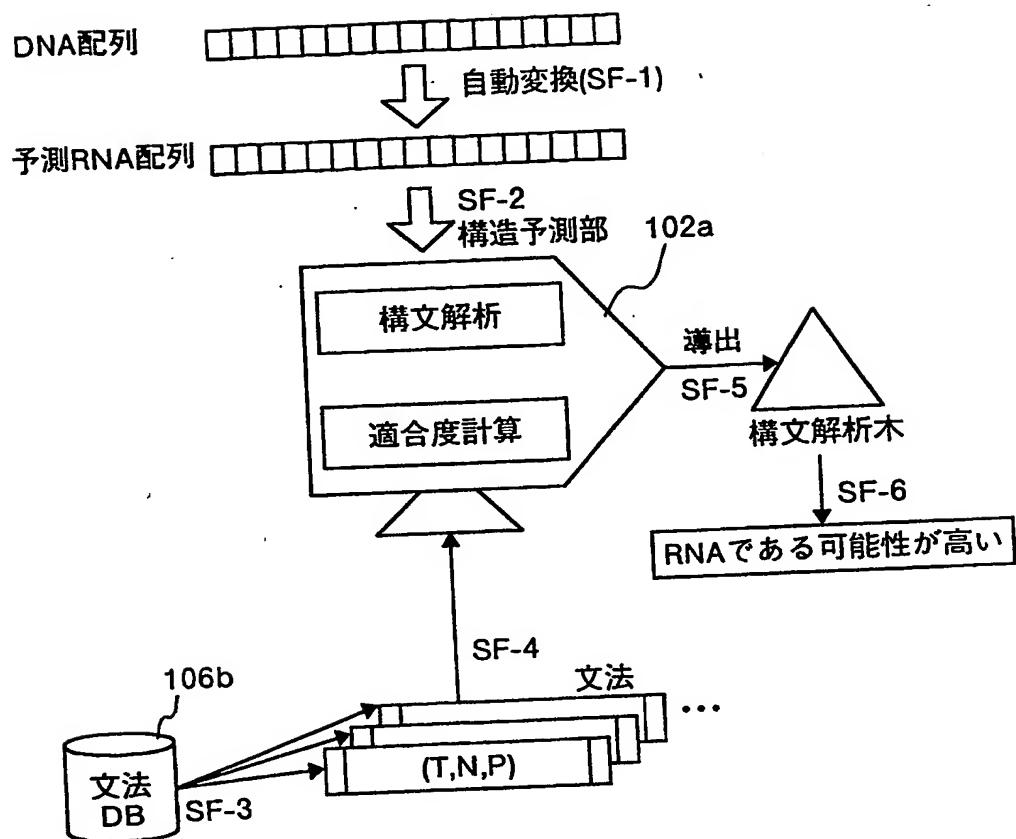
【図9】



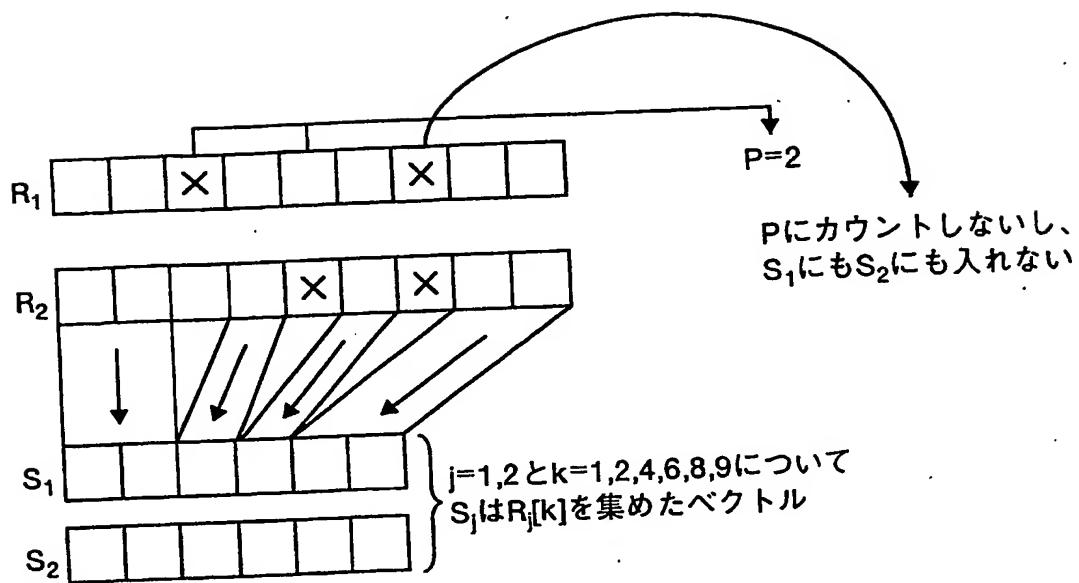
【図10】



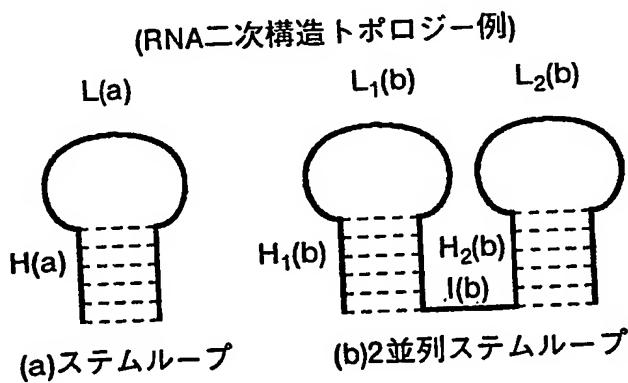
【図11】



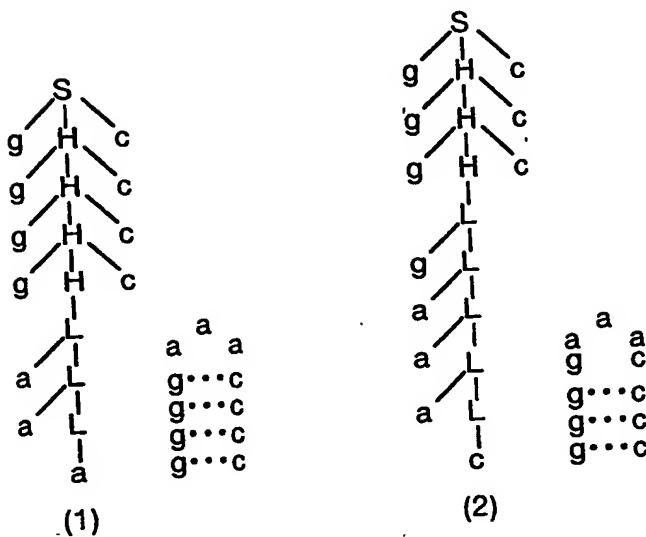
【図12】



【図13】



【図14】

(s₁の構文解析木と二次構造)

【図15】

(塩基対の自由エネルギー[kcal/mol])

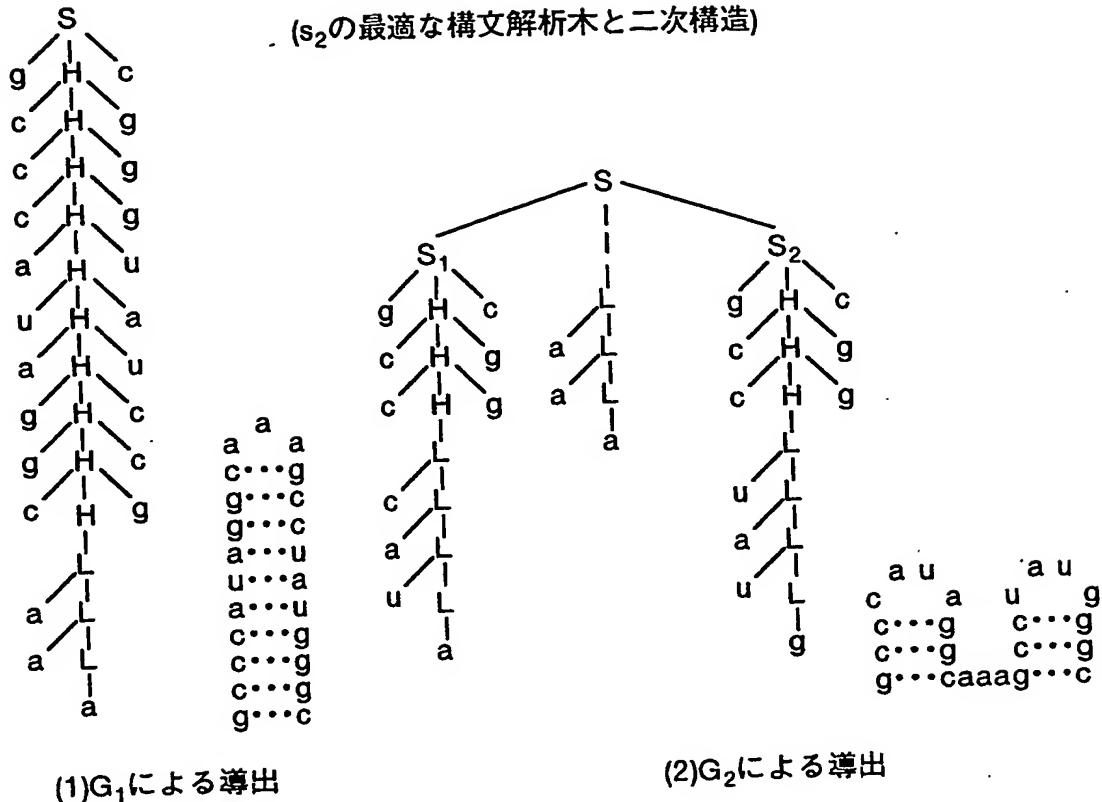
3'側塩基対	5'側塩基対				
	gu	au	ua	cg	gc
gu	-0.5	-0.5	-0.7	-1.5	-1.3
au	-0.5	-0.9	-1.1	-1.8	-2.3
ua	-0.7	-0.9	-0.9	-1.7	-2.1
cg	-1.9	-2.1	-2.3	-2.9	-3.4
gc	-1.5	-1.7	-1.8	-2.0	-2.9

【図16】

(ループの自由エネルギー[kcal/mol])

ループ種別	ループサイズ												
	1	2	3	4	5	6	7	8	9	10	12	14	16
バルジループ	3.3	5.2	6.0	6.7	7.4	8.2	9.1	10.0	10.5	11.0	11.8	12.5	13.0
バルビンループ	-	-	7.4	5.9	4.4	4.3	4.1	4.1	4.2	4.3	4.9	5.6	6.1
内部ループ	-	0.8	1.3	1.7	2.1	2.5	2.6	2.8	3.1	3.6	4.4	5.1	5.6

【図17】



【図18】

順位	トポロジー	適合度	二次構造
1	ステムループ T ₁	13.7	図18(1)に示された二次構造
2	2並列ステムループ T ₂	6.7	図18(2)に示された二次構造

【図19】

配列	適合度	二次構造
s_2	6.7	図18(2)に示された二次構造

【図20】

	T_1	T_2
s_1	1.4	-
s_2	13.6	6.7

【図21】

(sの最適な二次構造)

g g c a g
 g c
 a a c
 a a a
 a a a
 a c...g a
 c...g
 c...g
 g...c

(1)

a a a
 a a a
 c...g c...g
 c...g c...g
 c...g c...g
 g...c a g...c

(2)

【図22】

	T ₁	T ₂
s	-	5.1

2001-402081

【図23】

配列	適合度	トポロジー(スコア)
s_2	6.7	$T_2(5.1)$

【書類名】 要約書

【要約】

【課題】 RNA配列や抽出した文法を統合的に管理し、集積した文法やRNA配列を用いて二次構造予測などをより効率的に行うことのできるRNA配列解析装置等を提供することを課題とする。

【解決手段】 本発明にかかるシステムは、配列情報を解析するRNA配列解析装置であるRNA配列解析装置100と、配列情報等に関する外部データベースやホモロジー検索用の外部分析プログラム等を提供する外部システム200とを、ネットワーク300を介して通信可能に接続して構成されている。

【選択図】 図5

出願人履歴情報

識別番号 [500520628]

1. 変更年月日 2000年10月26日

[変更理由] 新規登録

住 所 千葉県千葉市美浜区中瀬1丁目3番地 幕張テクノガーデンD

17

氏 名 セレスター・レキシコ・サイエンシズ株式会社